Université du Maine

LES COLLECTIONS VOLUMINEUSES DE DOCUMENTS AUDIOVISUELS : SEGMENTATION ET REGROUPEMENT EN LOCUTEURS

Thèse

présentée et soutenue publiquement le
vendredi 3 juillet 2015
pour l'obtention du

Doctorat de l'Université du Maine

(spécialité informatique - ED STIM 503)
par

Grégor Dupuy

Rapporteurs

M. Claude Barras	Maître de Conférences, HDR	LIMSI-CNRS, Orsay
M. Nicholas Evans	Maître de Conférences, HDR	EURECOM, Biot

Examinateurs

M. Paul Deléglise	Professeur des Universités	LIUM, Le Mans
Mme Corinne Fredouille	Maître de Conférences	CERI-LIA, Avignon
M. Guillaume Gravier	Directeur de Recherche, CNRS	IRISA, Rennes
M. Denis Jouvet	Directeur de Recherche, INRIA	LORIA, Villers-lès-Nancy

Encadrants

M. Yannick Estève	Professeur des Universités	LIUM, Le Mans
M. Sylvain Meignier	Maître de Conférences, HDR	LIUM, Le Mans

Laboratoire d'Informatique de l'Université du Maine

lium

Remerciements

Résumé

La tâche de Segmentation et Regroupement en Locuteurs (SRL), telle que définie par le NIST, considère le traitement des enregistrements d'un corpus comme des problèmes indépendants. Les enregistrements sont traités séparément, et le taux d'erreur global sur le corpus correspond finalement à une moyenne pondérée. Dans ce contexte, les locuteurs détectés par le système sont identifiés par des étiquettes anonymes propres à chaque enregistrement. Un même locuteur qui interviendrait dans plusieurs enregistrements sera donc identifié par des étiquettes différentes selon les enregistrements. Cette situation est pourtant très fréquente dans les émissions journalistiques d'information : les présentateurs, les journalistes et autres invités qui animent une émission interviennent généralement de manière récurrente.

En conséquence, la tâche de SRL a depuis peu été considérée dans un contexte plus large, où les locuteurs récurrents doivent être identifiés de manière unique dans tous les enregistrements qui composent un corpus. Cette généralisation du problème de regroupement en locuteurs va de pair avec l'émergence du concept de collection, qui se réfère, dans le cadre de la SRL, à un ensemble d'enregistrements ayant une ou plusieurs caractéristiques communes.

Le travail proposé dans cette thèse concerne le regroupement en locuteurs sur des collections de documents audiovisuels volumineuses (plusieurs dizaines d'heures d'enregistrements). L'objectif principal est de proposer (ou adapter) des approches de regroupement afin de traiter efficacement de gros volumes de données, tout en détectant les locuteurs récurrents. L'efficacité des approches proposées est étudiée sous deux aspects : d'une part, la qualité des segmentations produites (en termes de taux d'erreur), et d'autre part, la durée nécessaire pour effectuer les traitements. Nous proposons à cet effet deux architectures adaptées au regroupement en locuteurs sur des collections de documents. Nous proposons une approche de simplification où le problème de regroupement est représenté par une graphe non-orienté. La décomposition de ce graphe en composantes connexes permet de décomposer le problème de regroupement en un certain nombre de sous-problèmes indépendants. La résolution de ces sous-problèmes de regroupement est expérimentée avec deux approches

de regroupements différentes (HAC et ILP) tirant parti des récentes avancées en modélisation du locuteur (i-vector et PLDA).

Abstract

The task of speaker diarization, as defined by NIST, considers the recordings from a corpus as independent processes. The recordings are processed separately, and the overall error rate is a weighted average. In this context, detected speakers are identified by anonymous labels specific to each recording. Therefore, a speaker appearing in several recordings will be identified by a different label in each of the recordings. Yet, this situation is very common in broadcast news data: hosts, journalists and other guests may appear recurrently.

Consequently, speaker diarization has been recently considered in a broader context, where recurring speakers must be uniquely identified in every recording that compose a corpus. This generalization of the speaker partitioning problem goes hand in hand with the emergence of the concept of collections, which refers, in the context of speaker diarization, to a set of recordings sharing one or more common characteristics.

The work proposed in this thesis concerns speaker clustering of large audiovisual collections (several tens of hours of recordings). The main objective is to propose (or adapt) clustering approaches in order to efficiently process large volumes of data, while detecting recurrent speakers. The effectiveness of the proposed approaches is discussed from two point of view: first, the quality of the produced clustering (in terms of error rate), and secondly, the time required to perform the process. For this purpose, we propose two architectures designed to perform cross-show speaker diarization with collections of recordings. We propose a simplifying approach to decomposing a large clustering problem in several independent sub-problems. Solving these sub-problems is done with either of two clustering approaches which take advantage of the recent advances in speaker modeling.

Sommaire

CHAPITRE 1

Introduction

Ces dernières années ont été marquées par l'évolution des technologies du multimédia, qui se sont multipliées, diversifiées et démocratisées : ordinateurs portables, appareils photographiques numériques, téléphones portables *(smartphones)*, tablettes tactiles, caméras vidéo *(GoPro)*, accessoires *connectés* (montres, bracelets, lunettes), *etc*. Les récentes avancées techniques et technologiques permettent à chacun d'enregistrer et de visualiser, à tout moment, des documents audiovisuels en (très) haute définition. L'essor de la télévision et d'Internet, avec ses plateformes de partage en ligne et ses réseaux sociaux, ainsi que l'amélioration des capacités de stockage en ligne et des infrastructures réseaux, ont fait que la quantité de données multimédia accessible ne cesse d'augmenter. Début juin 2014, YouTube annonçait [1] la mise en ligne de 100 heures de vidéo chaque minute. Le 15 février 2015, ce n'est plus 100, mais 300 heures de vidéo qui sont publiées chaque minute. Ce qui reviendrait à mettre en ligne, chaque minute, une vidéo dont la durée totale correspondrait à 12,5 jours. C'est donc l'équivalent de 18000 jours (soit environ 49 ans...) de contenu vidéo qui est publié, chaque jour, sur YouTube. L'Institut National de l'Audiovisuel recense quant à lui plus de 5 millions d'heures d'enregistrements [2] vidéos et radios à la fin de l'année 2013.

La plupart des travaux de recherche menés dans le domaine du traitement automatique de la parole ont porté sur l'analyse isolée de documents audio : les différents enregistrements qui composent un corpus de données sont traités indépendamment les uns des autres. Compte tenu des récentes avancées technologiques dans le domaine du multimédia, il apparaît intéressant de considérer les travaux de recherche

1. https://www.youtube.com/yt/press/fr/statistics.html
2. http://www.institut-national-audiovisuel.fr/nous-connaitre/entreprise/chiffres-cles.html

dans un contexte plus large, où les documents audiovisuels seraient regroupés en *collections*. Dans ce chapitre d'introduction, nous allons en premier lieu définir le concept de collection et présenter les pistes d'analyse envisagées lors de la rédaction de ce mémoire de thèse. Nous introduirons ainsi la tâche de Segmentation et Regroupement en Locuteur (SRL), en discutant à la fois sur les conséquences induites par le changement d'échelle (collection *vs.* émission), ainsi que sur les contraintes liées au contexte d'utilisation (évaluation *vs.* application).

1.1. Segmentation et Regroupement en Locuteurs

La tâche de Segmentation et Regroupement en Locuteurs (SRL), plus connue sous la dénomination *« Speaker Diarization »*, a été formellement décrite par le *National Institute of Standards and Technology (NIST)*[3] lors des campagnes d'évaluations *Rich Transcription (RT)* [NIST, 2003]. Le principal objectif des campagnes RT était d'enrichir les transcriptions des systèmes de reconnaissance automatique de la parole par des métadonnées, de manière à les rendre plus lisibles. La SRL, qui vise à répondre à la question « Qui parle quand ? », a été définie comme le découpage d'un flux audio en segments homogènes en fonction de l'identité des locuteurs. Il s'agit de détecter les tours de parole, qui correspondent aux changements de locuteurs, et d'identifier les segments de parole correspondant à un même locuteur par une étiquette unique au sein d'un document audio. Cette thématique de recherche s'inscrit dans domaine plus vaste de la reconnaissance automatique du locuteur (RAL), au même titre que la vérification et l'identification du locuteur. En revanche, la SRL ne dispose d'aucune information *a priori* sur les locuteurs. Historiquement, les premiers travaux effectués en SRL remontent au début des années 1990. Ils sont attribués à la société *BBN* qui cherchait à indexer automatiquement les enregistrements de conversations entre pilotes de ligne et contrôleurs aériens [Gish et al., 1991; Siu et al., 1992].

Trois principaux domaines d'application ont été particulièrement étudiés dans un contexte de recherche en SRL [Reynolds et Torres-Carrasquillo, 2005] :

- Les émissions journalistiques d'information *(broadcast news)*, dont les enregistrements étudiés correspondent à des émissions télévisuelles ou radiophoniques de type « journal d'information ». Ces émissions sont généralement caractérisées par l'intervention de plusieurs locuteurs et par une variabilité acoustique importante liée aux conditions d'enregistrement (studio, téléphone, reportage en extérieur, environnement bruité, présence de musique, publicités commerciales, etc.).

3. http://http://www.nist.gov/

- Les conversations téléphoniques, dont les enregistrements correspondent à des conversations orales entre deux ou plusieurs locuteurs, par l'intermédiaire d'un téléphone. Ce domaine d'application est essentiellement étudié dans le cadre de la vérification du locuteur.
- Les enregistrements de réunions *(meeting)*, qui sont principalement caractérisés par la présence de plusieurs locuteurs, pouvant communiquer depuis différents lieux, au moyen de plusieurs microphones. Le phénomène de parole superposée et la variabilité du canal sont particulièrement présents dans ce type d'enregistrements.

Avec la tâche de SRL, telle que définie par NIST, les enregistrements d'un corpus sont traités séparément. Le taux d'erreur global sur le corpus correspond finalement à une moyenne, pondérée par la durée des enregistrements traités, des taux d'erreur obtenus sur chaque enregistrement concerné. Dans ce contexte, les locuteurs détectés par le système sont identifiés par des étiquettes anonymes propres à chaque enregistrement. Un même locuteur qui interviendrait dans plusieurs enregistrements sera donc identifié par des étiquettes différentes selon les enregistrements. Cette situation est pourtant très fréquente dans les émissions journalistiques d'information. Les présentateurs, les journalistes et autres invités qui animent une émission interviennent généralement de manière récurrente d'un enregistrement à l'autre.

En conséquence, la tâche de Segmentation et Regroupement en Locuteurs a depuis peu été considérée dans un contexte plus large, où les locuteurs récurrents doivent être identifiés par une seule et même étiquette dans tous les enregistrements qui composent un corpus [Tran et al., 2011; Yang et al., 2011]. Cette nouvelle approche va de pair avec l'émergence du concept de *collection*, qui se réfère, dans le cadre de la SRL, à un ensemble d'enregistrements ayant une ou plusieurs caractéristiques communes. Dans ce manuscrit, nous désignons cette évolution naturelle de la tâche par **« SRL de collections »** *(cross-show speaker diarization)*. De manière contrastive, nous désignons par **« SRL d'émissions »** la tâche de SRL traitant les émissions d'un corpus, ou d'une collection, sans se préoccuper de la récurrence des locuteurs (*single-show speaker diarization*, ou simplement *speaker diarization*). La tâche de « SRL de collections » partage le même objectif que la tâche d'appariement en locuteurs *(speaker linking)* [Bourlard et al., 2013; Ferràs et Bourlard, 2012; Ghaemmaghami et al., 2013; Leeuwen, 2010; Meignier et al., 2002], c'est-à-dire, identifier les différentes interventions orales des locuteurs dans un ensemble d'enregistrements audio. Les approches mises en oeuvre sont similaires, la seule « différence » notable semble être que la tâche d'appariement en locuteurs est considérée comme une tâche annexe à la SRL, effectuée *a posteriori* à partir des segmentations produites par un système de SRL d'émissions. Dans le cadre de la SRL de collec-

tions, le procédé d'appariement en locuteurs est considéré comme la dernière étape de regroupement du système de SRL.

Les travaux menés dans le cadre de cette thèse sont exclusivement consacrés aux émissions journalistiques d'information. Dans les parties suivantes, nous présentons notre conception de la notion de collection, et posons les bases de la problématique sur laquelle a porté nos réflexions et nos travaux.

1.2. Définition du concept de collection

La littérature dans le domaine du traitement automatique de la parole définit le concept de collection comme un ensemble de documents audiovisuels ayant une ou plusieurs caractéristiques communes. La notion de collection peut être envisagée sous différents aspects. Ainsi, différents enregistrements d'une même émission, sur une certaine durée, constituent une collection. Des enregistrements d'émissions différentes, couvrant un même évènement médiatique, constituent une collection. Un ensemble d'enregistrements dans lesquels interviendrait une personnalité en particulier constitue une collection. Il convient donc de définir ce à quoi correspond une collection, ainsi que les différentes perspectives d'étude envisagées dans le cadre de cette thèse sur la tâche de segmentation et regroupement en locuteurs.

Les travaux réalisés dans le cadre de cette thèse concernent essentiellement les approches de traitement pour la SRL de collections, cependant, nous avons défini plusieurs genres de collections pour lesquelles la SRL de collections présenterait un intérêt notable à être étudiée (*cf.* schéma 1.1) :

Premièrement, les collections dites *« d'émissions »*, qui seraient composées d'enregistrements présentant des caractéristiques communes. Cette perspective est à prendre au sens large, les collections d'émissions étudiées dans cette thèse reposent sur trois niveaux de granularité différents : un niveau *Programme*, un niveau *Organisme* et un niveau *Thématique* (collections « horizontales » sur le schéma 1.1). Ensuite, les collections dites *« temporelles »*, qui seraient composées d'enregistrements de nature hétérogène ciblant une période temporelle précise (collections « verticales » sur le schéma 1.1). Nous pourrions également considérer des collections dont les enregistrements se rapporteraient à un même type (par exemple, les débats politiques). Nous considérons toutefois ce genre de collections (dénommées collections *« typologiques »*) comme un cas particulier des collections d'émissions.

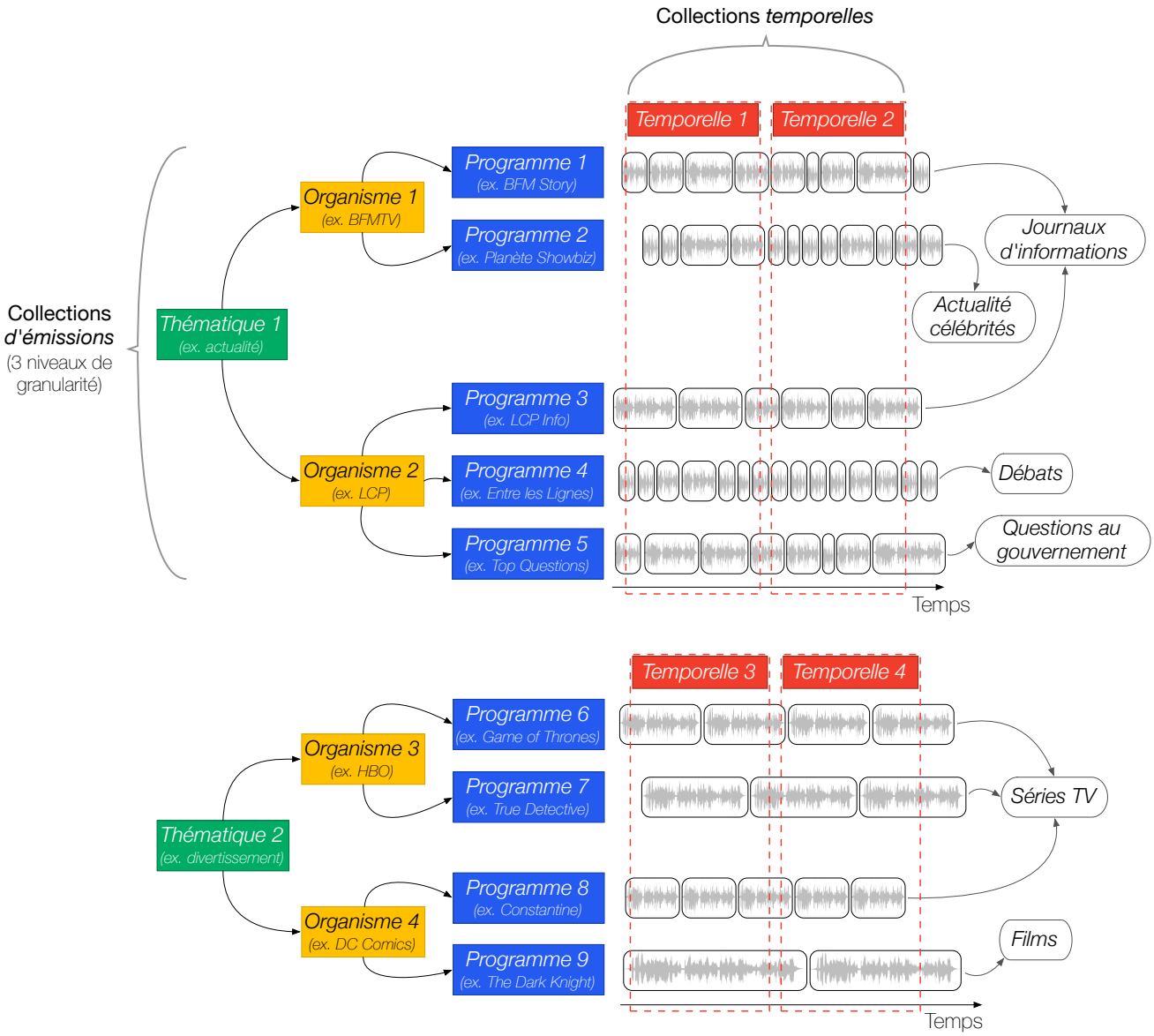

Figure *1.1 – Représentation schématique des perspectives étudiées : horizontalement, des collections d'émissions, et verticalement, des collections temporelles (le* type *des émissions est spécifié sur la partie droite).*

1.2.1 Collections d'émissions

Nous définissons les collections d'émissions comme un corpus d'enregistrements audiovisuels ayant une ou plusieurs caractéristiques communes. Dans cette thèse, nous avons considéré trois niveaux de granularité différents pour étudier les collections d'émissions, en fonction des enregistrements qui les constituent (*cf.* schéma 1.1):

1. Les collections d'émissions de niveau *Programme*, qui sont constituées d'enregistrements issus d'un même programme télévisuel. Une collection d'émissions de niveau *Programme*, par exemple, la collection *BFM Story*, serait constituée d'un ensemble d'enregistrements provenant exclusivement de l'émission *BFM Story*.
2. Les collections d'émissions de niveau *Organisme*, composées d'enregistrements provenant d'une même chaîne de télévision. Une collection d'émissions de niveau *Organisme*, par exemple, la collection *BFMTV*, serait composée d'enregistrements provenant exclusivement de la chaîne *BFMTV*, quel qu'en soit le programme télévisuel (inclusion de tous les enregistrements constituant les collections d'émissions de niveau *Programme* issus de la chaîne *BFMTV*, en l'occurrence, les collections *BFM Story* et *Planète Showbiz*).
3. Les collections d'émissions de niveau *Thématique*, qui regroupent tous les enregistrements d'émissions se rapportant à un même thème général. Une collection d'émissions de niveau *Thématique* dont le thème serait, par exemple, les émissions se rapportant à l'actualité, serait composée d'enregistrements provenant des chaînes *BFMTV* et *LCP*, tous programmes confondus (inclusion de tous les enregistrements constituant les collections d'émissions de niveau *Organisme*, en l'occurrence, les collections *BFMTV* et *LCP*).

La plupart des locuteurs récurrents correspondent alors aux personnes participantes et animant les émissions, telles que les présentateurs, les journalistes, les chroniqueurs, etc. L'étude des collections d'émissions peut, par exemple, mener à l'établissement d'un schéma d'interactions entre les locuteurs (« qui parle avec qui ? »), et permettre d'inférer le rôle des différents participants [Bigot et al., 2012].

1.2.2 Collections temporelles

Une collection temporelle est constituée d'enregistrements couvrant une période bien déterminée, ciblant par exemple un évènement de l'actualité hautement médiatisé. La nature des enregistrements n'aurait finalement que peu d'importance. Il

s'agirait dans l'idéal de regrouper des émissions associées à un même évènement médiatique, couvert sous différents angles par différents médias, et confronter ces documents afin d'en dégager une analyse croisée. L'étude des enregistrements de l'actualité sur une période temporelle bien déterminée permettrait de faire ressortir les principaux acteurs d'évènements médiatiques particuliers. Cette perspective applicative pourrait par exemple être envisagée dans le cadre d'une application de suivi de l'actualité.

1.2.3 Collections typologiques

Avec une collection typologique, la provenance des enregistrements aurait moins d'importance que l'objet ou l'intérêt des émissions. Il s'agirait de regrouper les enregistrements des émissions d'un même type, par exemple, les débats politiques. L'étude des collections typologiques permettrait de mettre en avant les principaux interlocuteurs d'un domaine en particulier.

Certains rapprochements peuvent être faits entre les collections typologiques et les collections d'émissions. En effet, les collections d'émissions de niveaux *Programme* et *Thématique* peuvent correspondre à des collections typologiques. Prenons par exemple les enregistrements de l'émission *Planète Showbiz*, de la chaîne *BFMTV*, pour illustrer ce propos. Les enregistrements de cette émission en particulier portent sur l'actualité des personnalités politiques. Parmi les données que nous utilisons dans le cadre de cette thèse, qui seront présentées ultérieurement dans ce manuscrit, il s'agit de la seule émission de ce type. Les enregistrements de l'émission *Culture et Vous* constituent donc à eux seuls une collection d'émissions de niveau *Programme* ainsi qu'une collection typologique (sur l'actualité des personnalités politiques). Les enregistrements de l'émission *Culture et Vous* s'intègrent également dans la collection d'émissions de niveau *Thématique* sur les émissions se rapportant à l'actualité.

1.2.4 Remarques

Des choix ont dû être faits quant à la manière de constituer les collections. Il est relativement simple de catégoriser les émissions comme *BFM Story* et *LCP Info*, qui correspondent à des journaux d'information. Il est, en revanche, plus difficile de se positionner sur les émissions comme *Planète Showbiz* (actualité des célébrités) et *Top Questions* (questions au gouvernement). Est-il raisonnable de considérer les enregistrements de ces émissions comme des enregistrements se rapportant à la thématique de l'actualité ? L'émission *Planète Showbiz* pourrait très bien être considérée

comme un divertissement. La classification des collections que nous proposons dans ce manuscrit repose sur des choix arbitraires et n'a pas pour vocation de définir une norme.

1.3. Positionnement du problème

En SRL d'émissions, nous traitons les différents enregistrements d'émissions journalistiques d'information séparément. En SRL de collections, nous souhaitons détecter les locuteurs récurrents dans un ensemble d'enregistrements. Le problème posé par la SRL de collections peut se résumer simplement à l'aide du schéma présenté en figure 1.2.

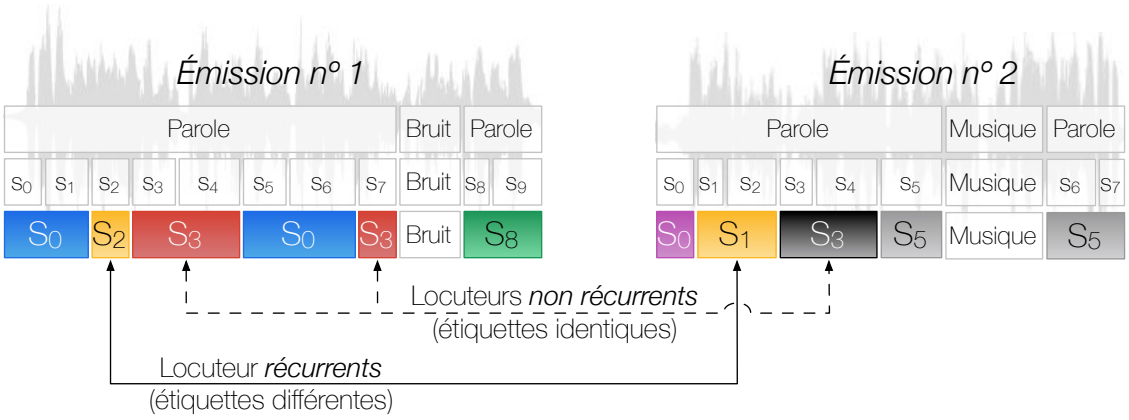

Figure *1.2 – Schématisation du problème de segmentation et regroupement en locuteurs dans le cadre du traitement de collections d'enregistrements.*

Sur cet exemple, nous simplifions le problème en considérant une collection composée de seulement deux émissions, dont les enregistrements auraient été segmentés et regroupés en locuteurs avec un système de SRL d'émissions. Sur ce schéma, les blocs colorés représentent les classes de locuteur obtenues à l'issue du procédé de SRL. Étant donné que le traitement des enregistrements est réalisé indépendamment, ces classes sont identifiées par une étiquette propre à l'enregistrement (S_0, S_2, *etc.*). Sur ce schéma, la couleur d'un bloc symbolise l'identité **réelle** du locuteur.

Le procédé de SRL de collection doit être en mesure de donner une étiquette unique à la collection pour chaque locuteur, de manière à ce que d'un enregistrement à l'autre, les locuteurs récurrents et non récurrents soient formellement identifiés. Or, à l'issue du procédé de SRL d'émission :

- Les locuteurs récurrents ne portent par la même étiquette d'une émission à l'autre (cas du locuteur symbolisé par le bloc jaune dans le schéma 1.2).

- Les locuteurs non récurrents peuvent être identifiés par la même étiquette d'une émission à l'autre (cas du locuteur étiqueté S_3 sur la segmentation des deux enregistrements du schéma 1.2)

L'approche la plus naïve, et cependant efficace, consiste à considérer les segmentations produites par le système de SRL d'émissions pour les différents enregistrements d'une collection et effectuer une nouvelle étape de regroupement. Le principal problème lié à ce changement d'échelle concerne la durée de traitement induite par la complexité de l'étape de regroupement. Cette durée est fonction de plusieurs facteurs : la quantité de données qui compose la collection, la complexité de l'algorithme de regroupement, la complexité des approches de modélisation du locuteur, et la puissance de calcul disponible. Les approches de regroupement et les techniques de modélisation à l'état de l'art utilisées en SRL d'émission permettent d'obtenir de très bons résultats, mais sont malheureusement très coûteuses en temps et en ressources.

Il s'agit là du principal problème auquel nous sommes confrontés. En effet, si une approche naïve permet d'obtenir des résultats satisfaisants sur des collections de petite taille, elle montre rapidement ses limites sur des collections plus volumineuses. Or, nous souhaitons traiter des collections dont le volume représente plusieurs dizaines d'heures d'enregistrements (la plus volumineuse des collections étudiées dans le cadre de cette thèse, présentée dans le chapitre 4, représente environ 178 heures d'audio). Il convient donc de proposer des approches adaptées, de manière à traiter des collections volumineuses en une durée de traitement raisonnable tout en produisant des segmentations de qualité.

1.4. Plan de ce manuscrit

Ce manuscrit de thèse est essentiellement composé de deux parties distinctes. Dans la première partie de ce manuscrit, composée des chapitres 2 et 3, nous présentons un état de l'art sur les deux composantes de notre problématique d'étude. Nous y présentons les approches, techniques et architectures principalement utilisées en SRL d'émissions et SRL de collections.

Une seconde partie, qui regroupe les chapitres 4, 5 et 6, présente les contributions réalisées dans le cadre de cette thèse. Nous présentons, dans le chapitre 4, les données audio et les différentes collections étudiées expérimentalement. Les chapitres 5 et 6 sont dédiés à la présentation des approches proposées. Nous présentons finalement nos conclusions et perspectives dans une troisième partie.

Nous proposons également une partie annexe organisée en 4 chapitres distincts. Les annexes A et B sont dédiées à la présentation des analyses intermédiaires sur lesquelles reposent les observations finalement établies quant aux approches proposées dans les chapitres 5 et 6. Les annexes C et D concernent quant à elles les études préliminaires sur lesquelles reposent les travaux présentés dans les chapitres 5 et 6.

Première partie

État de l'art

CHAPITRE 2

État de l'art en SRL d'émissions

Ce chapitre est dédié à la présentation des principales approches et techniques utilisées dans le cadre de la tâche de Segmentation et Regroupement en Locuteurs. Seules les techniques relatives à la tâche de SRL dans le contexte d'émissions journalistiques d'information seront abordées. Les techniques spécifiques aux autres domaines de SRL (conversations téléphoniques et enregistrements de réunions) n'entrent pas dans le cadre de cette thèse et ne sont donc pas présentés. Précisons également que ce qui suit concerne la SRL d'émissions. Les méthodes employées dans le cadre de la SRL de collections, qui sont complémentaires, sont présentées dans le chapitre suivant.

Nous proposons, en premier lieu, une présentation succincte de la tâche et de l'architecture générale d'un système de SRL d'émissions. Nous aborderons ensuite les approches mises en œuvre afin d'obtenir le résultat souhaité en nous appuyant sur le système élaboré au LIUM. Nous détaillerons ainsi les concepts de paramétrisation acoustique, segmentation, classification et regroupement en locuteur. Ces approches impliquent généralement une représentation statistique des locuteurs, dont les techniques de modélisation seront détaillées en fonction des particularités propres aux approches présentées. Dans cette partie de l'état de l'art dédiée à la SRL d'émission, l'accent porte essentiellement sur la dernière étape de regroupement en locuteurs, qui permet d'optimiser les segmentations produites pour les besoins de la tâche de SRL. Les étapes préalables visent à produire des segmentations où les classes de locuteurs sont les plus pures possible. Il s'agit, d'ailleurs, des segmentations utilisées pour la tâche de transcription automatique de la parole, car chaque classe est censée représenter la voix d'un seul locuteur. Ces étapes préalables ne sont en aucun cas facultatives, mais sont cependant en marge du travail effectué durant cette thèse. Nous distinguons donc deux *niveaux* dans l'architecture d'un système de SRL d'émis-

sions, qui feront l'objet d'une présentation séparée dans deux parties distinctes. Enfin, nous introduisons la métrique d'évaluation DER *(Diarization Error Rate)*, permettant d'évaluer la qualité des segmentations produites pour la tâche de SRL.

2.1. Présentation générale

2.1.1 La tâche de SRL

La tâche de Segmentation et Regroupement en Locuteurs (SRL) tente de répondre à la question « Qui parle quand ? » à tout moment d'un enregistrement audio. Concrètement, il s'agit de générer automatiquement un fichier d'indexation – ou « segmentation » –, dans lequel les segments du signal de la parole, caractérisés par un instant de début et une durée, sont associés à des locuteurs. L'une des applications de la SRL est la tâche de reconnaissance automatique de la parole (RAP). L'étape de segmentation, en séparant la parole des autres évènements acoustiques tels que les musiques, les jingles et les silences, permet aux systèmes de RAP de travailler exclusivement sur les segments de parole contenus dans le signal audio. L'étape de regroupement permet quant à elle de fournir suffisamment de données pour l'adaptation des modèles acoustiques en fonction du locuteur. La SRL procure donc un cadre de travail intéressant à de nombreux égards pour la RAP. La segmentation permet également d'améliorer la lisibilité des transcriptions, en les structurant en fonction des locuteurs et des tours de parole. L'architecture générale d'un système de SRL est traditionnellement composée de quatre *modules* qui, à partir d'un signal audio supposé contenir de la parole, permettent d'en générer une segmentation en locuteurs [Barras et al., 2006; Fredouille et Evans, 2008; Gupta et al., 2008; Meignier et Merlin, 2009; Tranter et Reynolds, 2004]. Ces quatre modules, schématisés en figure 2.1, sont :

- La paramétrisation, qui vise à segmenter le signal en trames et en extraire des paramètres acoustiques.
- La segmentation en locuteurs, qui cherche à détecter les ruptures acoustiques et regroupe les trames consécutives en *segments*, dont les frontières correspondent aux changements de locuteurs.
- La segmentation parole/non-parole, qui vise à retirer les segments correspondant à des évènements acoustiques différents de la parole.
- Le regroupement en locuteurs, ou classification, qui a pour objectif de regrouper les segments correspondant à un même locuteur dans une classe unique,

permettant ainsi d'identifier les différentes zones du signal où le locuteur intervient oralement.

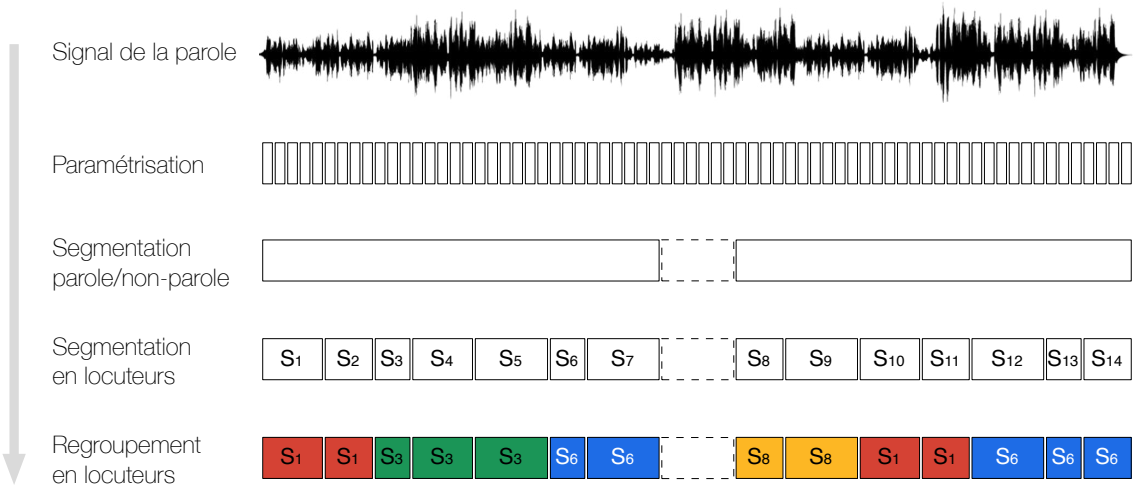

Figure *2.1 – Représentation schématique des quatre modules principaux de l'architecture d'un système de SRL d'émissions.*

Ce type d'architecture est généralement complétée par des modules supplémentaires, et l'ordre dans lequel interviennent les modules n'est pas figé. À titre d'exemple, l'architecture du système élaboré au LIUM intègre un module d'identification du genre et de la bande de fréquence. Dans ce même système, la segmentation parole/non-parole est réalisée en aval de la détection des ruptures [Meignier et Merlin, 2009], contrairement au système proposé par [Barras et al., 2006], où cette segmentation est réalisée immédiatement après l'étape de paramétrisation.

2.1.2 Architecture d'un système de SRL

Dans ce chapitre, nous proposons de nous appuyer sur l'architecture du système de SRL d'émission du LIUM pour présenter les approches et techniques mises en œuvre dans les différents modules évoqués précédemment. Cette démarche a également pour objectif d'introduire en détail le système de SRL d'émissions à partir duquel les travaux menés dans cette thèse ont été réalisés.

L'architecture du système élaboré au LIUM s'inspire de celle du système multi-niveau, décrite dans [Barras et al., 2006], ayant obtenu les meilleurs résultats en SRL d'émissions lors des campagnes d'évaluation Fall 2004 Rich Transcription [NIST, 2004] et ESTER [1] (2005) [Galliano et al., 2005]. Le système du LIUM, dont l'architecture est représentée en figure 2.2, a été initialement développé pour les tâches

1. Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques

de transcription automatique (campagne ESTER 1) et de SRL (campagne ESTER 2 [Galliano et al., 2009]). Ce système a permis au LIUM de se classer premier, ou second, dans les tâches de SRL des principales campagnes d'évaluation françaises sur des émissions journalistiques d'information, telles qu'ESTER 2 (2009), ETAPE (2011) [Gravier et al., 2012] et REPERE (janvier 2012, 2013 et 2014) [Galibert et Kahn, 2013]. L'outil *LIUM_SpkDiarization*, qui propose un ensemble de méthodes permettant d'élaborer des systèmes de SRL complets [Meignier et Merlin, 2009; Rouvier et al., 2013], exploite cette architecture par défaut. Cet outil de recherche est distribué [2] sous licence publique générale GNU (GPL).

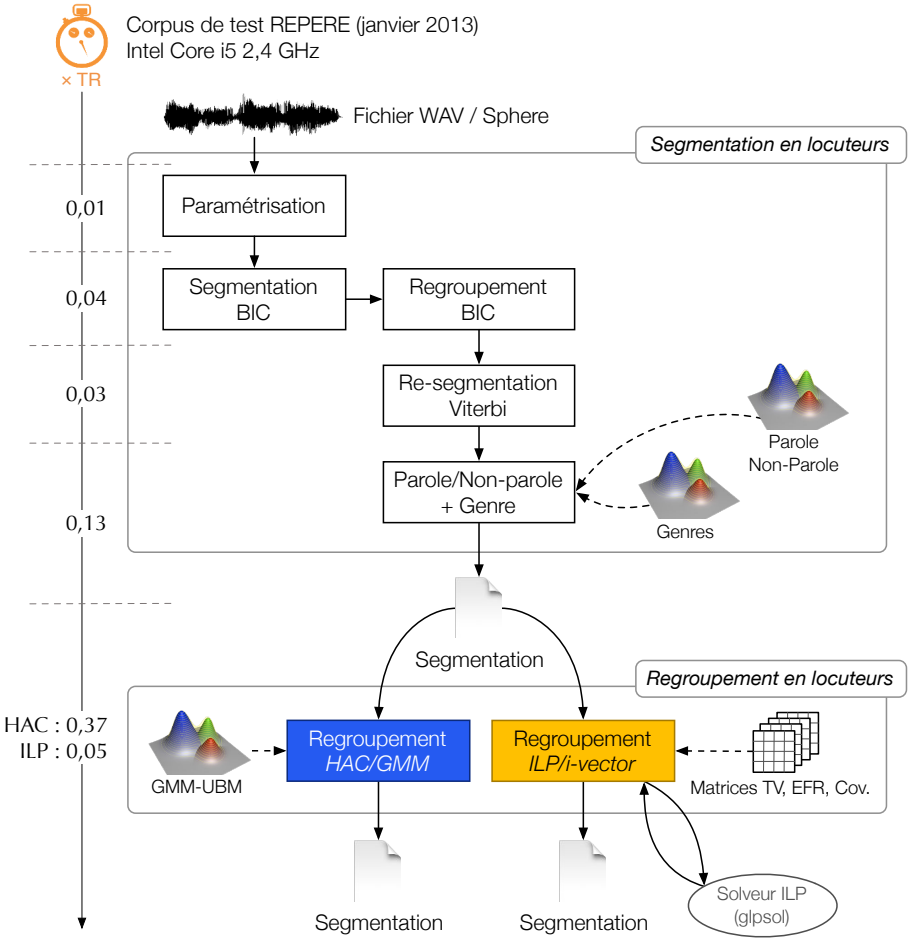

Figure *2.2 – Architecture du système de SRL d'émissions développé au LIUM. Les durées d'exécution de chaque étape sont exprimées en fraction du temps réel du corpus de test REPERE de janvier 2013.*

L'architecture de ce système peut être séparée en deux *composantes*. La première composante regroupe les approches visant à produire des classes de locuteurs les plus « pures » possible, où chaque classe est censée représenter un ensemble de segments de parole se rapportant à un seul locuteur (ces classes sont dépendantes du locuteur et des conditions acoustiques). Cette première composante apparaît, sur la figure 2.2,

2. http://www-lium.univ-lemans.fr/en/content/liumspkdiarization

sous la dénomination de « *Segmentation en locuteurs* » (bien que plusieurs phases de segmentation et regroupement soient alternées). La seconde composante concerne les approches mises en œuvre afin d'optimiser les segmentations pour satisfaire les objectifs de la tâche de SRL (« *Regroupement en locuteurs* » sur la figure 2.2) : les classes produites à l'issue de la première composante sont d'une grande pureté, cependant, plusieurs classes peuvent faire référence à un même locuteur, il est donc nécessaire de les regrouper.

Les travaux menés dans le cadre de cette thèse ne concernent essentiellement qu'un aspect isolé de la SRL d'émissions, le regroupement en locuteurs. Nous nous sommes donc limités, pour ce qui est de la description de la première composante de l'architecture, à ne présenter que les approches et techniques effectivement mises en place dans le cadre de nos travaux. En revanche, les approches et techniques présentées sont celles habituellement employées, à l'heure actuelle, pour la SRL d'émissions dans le contexte d'émissions journalistiques d'information [Barras et al., 2006; Gupta et al., 2008; Meignier et Merlin, 2009].

2.2. Segmentation en locuteurs (composante n°1)

Cette première composante de l'architecture alterne séquentiellement plusieurs phases de segmentation et de regroupement.

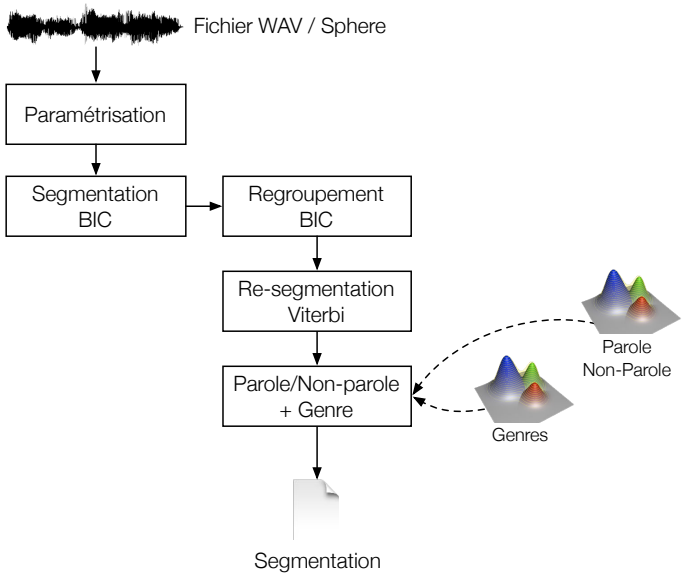

Figure *2.3 – Première composante de l'architecture du système de SRL d'émissions développé au LIUM, permettant la production de segmentations adaptées aux besoin de la transcription automatique.*

Elle prend le signal audio en entrée et produit une segmentation où le degré de

pureté des segments est très élevé (un segment ne « contient » la voix que d'un seul locuteur). Étant donné qu'il s'agit des segmentations utilisées pour la transcription automatique de la parole, ces segmentations sont également caractérisées par : la durée des segments, qui n'excède pas 20 secondes, et la frontière des segments, qui correspond tant que possible à des évènements acoustiques distincts de la parole (respiration, silence, *fillers*, . . .). Le cas échéant, les frontières sont positionnées entre deux mots pour minimiser les perturbations sur le modèle de langage.

2.2.1 Paramétrisation acoustique

La paramétrisation est un procédé consistant à représenter sous forme de vecteurs de coefficients acoustiques les informations jugées pertinentes du signal de la parole. Le signal de la parole est continu et non-stationnaire, il ne peut être utilisé tel quel du fait de sa variabilité et des limitations de représentation par les méthodes actuelles. Il peut néanmoins être considéré comme pseudo-stationnaire sur de très cours intervalles de temps (inférieurs à 100 millisecondes) [Rabiner et Juang, 1993]. La plupart des techniques d'analyse et de modélisation du signal de la parole reposent sur cette conjecture. Le signal est alors fragmenté en séquences, appelées *trames*, dont l'ordre de grandeur représente entre 20 et 50 millisecondes de signal.

La représentation cepstrale, qui présente l'avantage de séparer efficacement l'excitation glottique de la résonance induite par l'appareil vocal humain, est majoritairement utilisée dans les tâches de reconnaissance automatique de la parole (RAP), d'identification et vérification en locuteurs, et de SRL. Deux méthodes sont principalement utilisées pour encoder le signal en un jeu de coefficients : d'une part, les approches non-paramétriques, où le signal de la parole est représenté mathématiquement (transformée de Fourier à court terme, banc de filtres) [Oppenheim et Schafer, 1975; Rabiner et Schafer, 1978], d'autre part, les approches à base de modélisation paramétrique, où l'estimation des paramètres d'un modèle du signal de la parole permet de représenter convenablement les caractéristiques acoustiques de la parole (prédiction linéaire) [Atal et Hanauer, 1971; Makhoul, 1975]. Dans le cadre de nos travaux, nous n'avons eu recours qu'à la paramétrisation MFCC *(Mel-Frequency Cepstral Coefficients)*, dont les coefficients résultent d'un lissage des densités spectrales de chaque trame par les coefficients spectraux d'un banc de filtre triangulaire en échelle *Mel*.

▷ Normalisation des coefficients cepstraux

Le signal de la parole subit de nombreuses variations acoustiques, principalement liées aux conditions physiques et psychologiques du locuteur, à la qualité du matériel (microphones et canal de transmission) et aux conditions d'enregistrement (bruit, distorsions, réverbération, etc.). Ces informations *parasites* sont prises en compte dans la représentation cepstrale, elles introduisent un biais qu'il est nécessaire d'atténuer.

La technique de compensation la plus classique pour réduire les distorsions provoquées par le canal de transmission (les variations acoustiques engendrées par les locuteurs permettent au contraire de mieux les caractériser) consiste à centrer et réduire les coefficients par soustraction de leurs moyennes cepstrales (*Cepstral Mean Substraction* – CMS) [Furui, 1981]. Dans les travaux présentés dans ce manuscrit, nous lui préférons une variante où les coefficients sont, en plus, réduits par leurs variances (*Mean and Variance Normalization* – MVN). Nous appliquons également, en amont des techniques CMS et MVN, la méthode de normalisation *feature warping* [Ouellet et al., 2005; Pelecanos et Sridharan, 2001] (uniquement avec les approches de regroupement hiérarchique et de modélisation GMM, qui seront présentées en partie 2.3). Cette méthode de normalisation, complémentaire, permet de modifier la distribution des coefficients cepstraux en fonction d'une distribution normale cible.

▷ Enrichissement des trames

Les coefficients cepstraux permettant de représenter les trames du signal de la parole sont dits *statiques*. Ces coefficients statiques sont généralement enrichis par des informations *dynamiques* afin d'en apprécier la variation et l'évolution temporelle. Ces informations dynamiques, qui correspondent à la vitesse et à l'accélération des variations immédiates du spectre, sont estimées pour chaque trame à partir des dérivées temporelles première (Δ) et seconde ($\Delta\Delta$) de ses coefficients cepstraux [Furui, 1981]. L'énergie du signal et le premier coefficient cepstral C_0, ainsi que leurs dérivées premières et secondes, sont des coefficients discriminants pouvant également être incorporés aux vecteurs de paramètres modélisant les trames.

2.2.2 Segmentation BIC

Cette étape de segmentation, réalisée le plus souvent en deux temps en s'inspirant de l'approche proposée par [Delacourt et Wellekens, 2000], a pour but la production

de segments homogènes pouvant être exploités dans les étapes suivantes. Une première passe sur le signal est effectuée pour détecter les ruptures (changements de locuteurs), à l'aide de la mesure GLR (*Generalized Likelihood Ratio*). Une seconde passe permet, quant à elle, d'affiner la segmentation obtenue lors de la première passe en regroupant les segments consécutifs qui maximisent un score de vraisemblance : les segments étant moins nombreux et de longueur suffisante, il est possible d'utiliser une mesure plus discriminante (*Bayesian Information Criterion* – BIC).

▷ 1re passe : détection des ruptures

Cette première passe de segmentation permet au système de détecter les ruptures correspondant aux frontières des futurs segments. L'algorithme de segmentation [Siegler et al., 1997] repose sur le rapport de vraisemblance généralisé (*Generalized Likelihood Ratio* – GLR) et sur deux fenêtres glissantes temporelles consécutives qui parcourent l'intégralité du signal en se décalant d'un pas défini *a priori* (*cf.* figure 2.4).

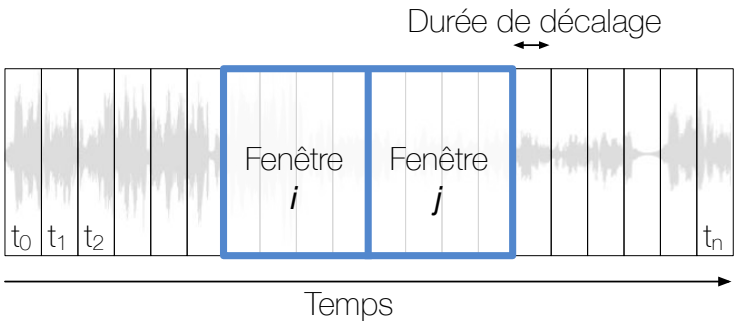

Figure *2.4 – Détection des ruptures acoustiques par mesure de dissimilarité entre les trames délimitées par les fenêtres glissantes adjacentes i et j.*

Les trames délimitées par ces fenêtres sont représentées par des modèles mono-gaussiens à matrices de covariance pleine. Le rapport de vraisemblance généralisé, qui est une mesure de distance fréquemment utilisée en SRL pour détecter les changements de locuteur, est alors déterminé à partir des deux modèles mono-gaussiens. Les ruptures, ou changements de locuteurs, correspondent alors aux valeurs maximales des mesures GLR obtenues. Cette métrique, introduite par [Gish et al., 1991], est un rapport de vraisemblance entre deux hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$:

- $\mathcal{H}_0$: les deux séquences x_i et x_j sont produites par un même locuteur x, auquel cas le modèle $M(\mu, \Sigma)$ correspondant à $x = x_i \cup x_j$ permettrait une meilleure représentation de x_i et x_j.
- $\mathcal{H}_1$: les deux séquences x_i et x_j sont produites par deux locuteurs différents, auquel cas les deux modèles $M_i(\mu_i, \Sigma_i)$ et $M_j(\mu_j, \Sigma_j)$ seraient plus adaptés pour

représenter x_i et x_j.

Le test de vraisemblance est ainsi formulé par le rapport des deux hypothèses :

$$GLR(x_i, x_j) = \frac{L(x, M(\mu, \Sigma))}{L(x_i, M_i(\mu_i, \Sigma_i))L(x_j, M_j(\mu_j, \Sigma_j))} \tag{2.1}$$

où $L(x, M(\mu, \Sigma))$ correspond à la vraisemblance de la séquence $x = x_i \cup x_j$ étant donné le modèle $M(\mu, \Sigma)$, et $L(x_i, M_i(\mu_i, \Sigma_i))L(x_j, M_j(\mu_j, \Sigma_j))$ la vraisemblance que les séquences x_i et x_j aient été produite par deux locuteurs différents.

La distance entre les deux séquences x_i et x_j est finalement établie par le logarithme du rapport de vraisemblance généralisé : $d(x_i, x_j) = -\log GLR(x_i, x_j)$. Le rapport de vraisemblance généralisé correspond au score de vraisemblance le plus fréquemment utilisé pour cette première passe, cependant, d'autres scores peuvent être employé, comme la version symétrique de la divergence de Kullback-Leibler (KL2) [Delacourt et al., 1999; Siegler et al., 1997], ou encore la mesure de divergence gaussienne (GD) [Barras et al., 2006].

Les ruptures détectées par la mesure GLR permettent de délimiter des segments dans le signal de la parole. Ces segments, constitués de trames consécutives, remplissent les conditions nécessaires à l'utilisation du critère d'information bayésien, réputé pour son fort pouvoir de différentiation entre les locuteurs.

▷ 2e passe : regroupement des segments consécutifs

À l'issue de ce processus est réalisée une deuxième passe visant à agglomérer les segments consécutifs. Les segments déterminés par les ruptures détectées lors de la première passe sont toujours représentés par des modèles mono-gaussiens à matrice de covariance pleine, cependant, la similarité entre ces modèles est désormais estimée par le critère d'information bayésien (BIC). La méthode du Critère d'Information Bayésien (*Bayesian Information Criterion* – BIC), introduite par [Schwarz et al., 1978] et employée en détection du locuteur par [Chen et Gopalakrishnan, 1998], est une métrique très appréciée pour sa simplicité et son efficacité. Cette métrique présente cependant un inconvénient : les segments de parole doivent être suffisamment longs sans quoi la quantité de données n'est pas suffisante pour estimer des modèles mono-gaussiens robustes. Or, grâce à la première passe de segmentation, il est désormais possible d'en tirer profit.

Étant donné un ensemble de modèles M, le critère BIC vise à sélectionner le modèle M_i le plus proche d'une distribution x constituée par n trames. Cette sélection

est réalisée par maximum de vraisemblance entre la distribution x et les modèles de l'ensemble M. La fonction de vraisemblance est pénalisée de manière a mieux estimer le degré de correspondance entre la distribution et les modèles, en tenant compte de la complexité de ces derniers :

$$BIC(M_i) = \log L(x, M_i) - \lambda \frac{1}{2} \#(M_i) \log(n) \tag{2.2}$$

où $L(x, M_i)$ correspond au maximum de vraisemblance entre les données de x et le modèle M_i, λ est un facteur de pénalité déterminé expérimentalement, $\#(M_i)$ est la complexité du modèle M_i.

La proximité de deux observations x_i et x_j peut être évaluée en réalisant la différence de leurs critères BIC respectifs (notée ΔBIC). À la manière du rapport GLR, il est possible de déterminer en fonction de la valeur obtenue si les données modélisées proviennent d'un même locuteur ou deux locuteurs différents : une valeur ΔBIC négative sous-entend que les distributions x_i et x_j seraient mieux représentées par deux modèles distincts.

$$\Delta BIC(x_i, x_j) = \frac{n_i + n_j}{2} \log |\Sigma x| - \frac{n_i}{2} \log |\Sigma x_i| - \frac{n_j}{2} \log |\Sigma x_j| + \lambda P \tag{2.3}$$

où $|\Sigma x_i|$, $|\Sigma x_j|$ et $|\Sigma x|$ désignent les déterminants des modèles mono-gaussiens représentants respectivement x_i, x_j, et $x = x_i \cup x_j$. λ est un paramètre à déterminer expérimentalement. P est le facteur de pénalité dépendant de la dimension des caractéristiques acoustiques de n_i et de n_j. Dans le cas de modèles mono-gaussiens à matrices de covariance pleines, le facteur de pénalité P est déterminé par la relation suivante, dans laquelle d représente la dimension des caractéristiques acoustiques :

$$P = \frac{1}{2} \left(d + \frac{d(d+1)}{2} \right) + \log (n_i + n_j) \tag{2.4}$$

Étant donné un ensemble de segments $\mathbf{S} = \{s_i, s_{i+1} \ \ldots, \ s_n\}$, l'algorithme consiste donc à regrouper les segments consécutifs s_i et s_{i+1} tant que la valeur du score ΔBIC entre leurs modèles mono-gaussiens respectifs est positive. Si la valeur du score ΔBIC est négative, alors l'algorithme passe aux segments suivants, et ainsi de suite jusqu'à s_n. Cette seconde passe permet finalement de constituer des segments suffisamment longs et purs pour améliorer la modélisation en locuteurs dans l'étape suivante.

2.2.3 Regroupement BIC

Nous considérons cette étape comme le premier regroupement en locuteurs de l'architecture. Contrairement à la deuxième passe de l'étape précédente, qui n'avait pour but que de constituer des segments suffisamment longs en *fusionnant* les segments consécutifs satisfaisant le critère BIC, nous cherchons ici à regrouper tous les segments correspondant à un même locuteur au sein d'une même classe.

L'algorithme mis en place correspond à un regroupement agglomératif hiérarchique (approche ascendante), pour lequel l'ensemble initial de classes est composé d'un unique segment par classe. Chacune des classes est représentée par un modèle mono-gaussien à matrice de covariance pleine, appris sur les données du segment de la classe, et le score de vraisemblance employé pour sélectionner les classes à regrouper est, à nouveau, le score ΔBIC (*cf.* équation 2.3). Cet algorithme consiste à regrouper itérativement les deux classes c_i et c_j les plus *proches*, c'est-à-dire, celles maximisant la valeur du score $\Delta BIC(c_i, c_j)$ calculé à partir des segments des classes c_i et c_j. Le procédé se poursuit tant que la valeur de la mesure $\Delta BIC_{c_i,c_j}$ est positive. Dans le cadre des émissions journalistiques d'information, nous fixons généralement le paramètre λ, qui influe sur le facteur de pénalité P (*cf.* équation 2.3), à 3.

L'approche de regroupement hiérarchique est présentée en détail dans la partie 2.3, qui est dédiée à la deuxième composante de notre architecture pour la SRL d'émissions, visant à optimiser les segmentations pour la tâche de SRL.

2.2.4 Re-segmentation par décodage de Viterbi

La segmentation issue du regroupement hiérarchique BIC est remise en question par un décodage de Viterbi, où chacune des classes résultant du regroupement BIC est modélisée par un modèle de Markov caché (*Hidden Markov Model* – HMM) [Baum et Petrie, 1966] mono-état. Les modèles de Markov cachés sont des automates probabilistes à états finis permettant de calculer la probabilité d'émission d'une séquence d'observations. Cette approche de modélisation acoustique, incontournable dans le domaine de la reconnaissance automatique de la parole [Baker, 1975; Jelinek, 1976; Rabiner, 1989], est également très appréciée en SRL : les HMM sont fréquemment utilisés pour segmenter le signal d'un enregistrement audio en fonction de ses caractéristiques acoustiques [Barras et al., 2004; Meignier et al., 2001], par l'intermédiaire d'un décodage de Viterbi [Viterbi, 1967].

Chacun des HMM mono-état est représenté par un modèle de mélanges gaussiens (*Gaussian Mixture Model* – GMM) à 8 composantes, appris via l'algorithme

espérance-maximisation (*Expectation-Maximisation* – EM), sur l'ensemble des segments de la classe. L'approche de modélisation GMM est devenue une référence en matière de modélisation du locuteur, et ce depuis son introduction en reconnaissance de la parole [Reynolds, 1992; Rose et Reynolds, 1990]. La modélisation GMM permet une estimation plus robuste des distributions des vecteurs acoustiques, et donc, une meilleure modélisation que les modèles mono-gaussiens. L'approche de modélisation GMM et ses deux principales méthodes d'apprentissage (EM et *Maximum A Posteriori*) seront présentées en détail dans la partie 2.3 dédiée à la deuxième composante de notre architecture pour la SRL d'émissions.

Ce décodage de Viterbi vise à produire une segmentation optimisée pour la reconnaissance de la parole. En effet, les frontières des segments produits jusqu'alors peuvent tomber entre deux phonèmes d'un même mot, ce qui pose problème pour générer des transcriptions robustes, car les propositions du modèle de langage peuvent être perturbées. Les frontières des segments produits sont donc légèrement déplacées pour correspondre aux zones de plus faible énergie du signal de la parole. De plus, les longs segments sont récursivement découpés de manière à ne pas excéder une durée de 20 secondes. Le déplacement des frontières et le découpage récursif sont réalisés par un ensemble de règles spécifiques, déterminées expérimentalement.

2.2.5 Détection parole/non-parole

L'étape de détection parole/non-parole vise à identifier la nature des caractéristiques acoustiques des segments, afin d'isoler, en particulier, les segments de parole. Il s'agit d'un second décodage de Viterbi, effectué avec un HMM composé de 8 mono-états. Chaque mono-état est représenté par un modèle GMM de 64 composantes gaussiennes, appris via l'algorithme EM sur les données d'apprentissage de la campagne d'évaluation ESTER 1. La paramétrisation employée pour l'apprentissage de ces 8 modèles GMM correspond à 12 paramètres MFCC augmentés de leurs coefficients différentiels Δ respectifs. La distinction est faite entre les deux bandes de fréquence habituellement utilisées en RAP pour les enregistrements studio et téléphoniques. Ces 8 modèles correspondent à : deux modèles de silence (un modèle studio et un modèle téléphone), trois modèles de parole studio (parole pure, parole sur du bruit et parole sur de la musique), un modèle de parole téléphone, un modèle de musique, et un modèle de *jingles*.

2.2.6 Détection du genre et de la bande de fréquence

La nature des segments étant désormais identifiée, il convient d'annoter les segments de parole en fonction du genre du locuteur (homme/femme) et du type de bande de fréquence (studio/téléphone). Ce procédé est réalisé au moyen de quatre modèles GMM composés de 128 composantes gaussiennes (un modèle par combinaison genre/bande de fréquence), appris sur environ une heure de parole issue des données d'apprentissage de la campagne ESTER 1, via la méthode EM. La paramétrisation employée pour créer ces quatre modèles GMM diffère légèrement de celle employée précédemment : les 12 paramètres MFCC + Δ respectifs sont normalisés par les méthodes de *feature warping* [Pelecanos et Sridharan, 2001] et MVN. Chaque segment est annoté en fonction du GMM maximisant le score de vraisemblance par rapport aux données du segment.

2.2.7 Bilan

La segmentation produite à l'issue de cette étape de détection du genre et de la bande de fréquence, et donc à l'issue de la première composante de notre architecture pour la SRL d'émissions, est adaptée aux besoins de la tâche de transcription automatique de la parole, qui adapte ses modèles acoustiques à la fois au locuteur et aux conditions acoustiques :

- Les segments de parole sont identifiés et annotés en fonction du genre et de la bande.
- La durée des segments n'excède pas 20 secondes.
- Un segment ne contient théoriquement que la voix d'un seul locuteur, pour des conditions acoustiques données (la contribution du canal est utilisée pour faire la distinction entre les locuteurs).

Ce système de SRL d'émissions, mis en place pour la campagne d'évaluation ESTER 2, a permis une diminution du taux d'erreur mot (WER) de 0,8% sur les transcriptions produites lors de la campagne d'évaluation ESTER 1, comparé au système utilisé à cette époque [Meignier et Merlin, 2009].

2.3. Regroupement en locuteurs (composante n°2)

Les segmentations produites par la première composante de notre système de SRL est tâche de transcription automatique ne satisfait pas les objectifs énoncés pour la

tâche de SRL, où le nombre final de partitions est idéalement censé correspondre au nombre de locuteurs présents dans l'enregistrement. À ce niveau, les classes de locuteurs finalement proposées sont supposées représenter des segments de parole énoncés par un seul et même locuteur. La pureté des classes est donc très élevée, cependant, plusieurs classes peuvent encore représenter un même locuteur. Il est donc de rigueur d'effectuer une nouvelle étape de regroupement en locuteurs en minimisant cette fois la contribution du canal, de manière à regrouper les classes correspondant à un même locuteur au sein d'une classe unique à l'enregistrement audio.

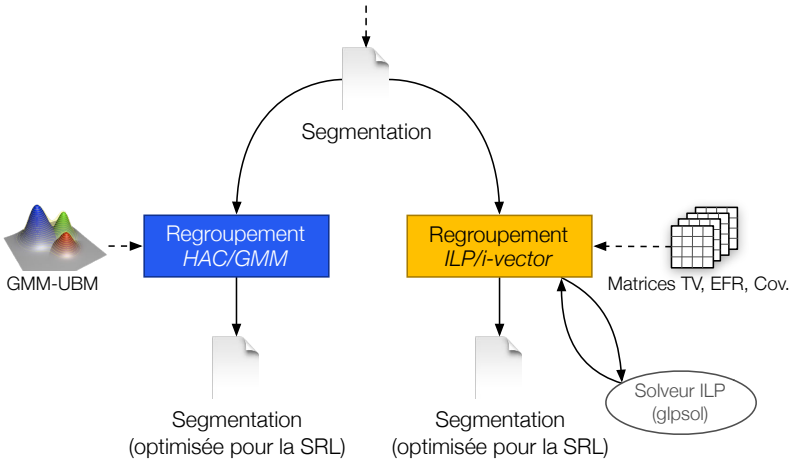

Figure *2.5 – Seconde composante de l'architecture du système de SRL d'émissions développé au LIUM, visant à produire des segmentations optimisées pour la SRL.*

Cette nouvelle étape de classification automatique est, là encore, totalement non supervisée. Elle est menée sans aucune information *a priori* sur les locuteurs : le nombre de locuteurs ainsi que leurs identités sont inconnus, et aucun modèle de leurs voix n'est disponible.

Par rapport à cette seconde composante de l'architecture, nous présentons tout d'abord les approches de modélisation GMM et i-vectors, qui permettent de modéliser les locuteurs de manière plus complexe et plus robuste que les modèles mono-gaussiens. Cette optimisation est permise en raison des segmentations fournies par la première composante du système de SRL, grâce auxquelles nous disposons de nombreux segments de parole pour modéliser une classe de locuteur. S'ensuit alors une présentation des principaux scores de vraisemblance employés pour estimer la proximité entre deux modèles GMM ou deux modèles i-vector. Enfin, nous présentons les deux méthodes de regroupement – ou classification – en locuteurs les plus étudiées au cours de cette thèse : le regroupement agglomératif hiérarchique, où les classes de locuteurs sont représentées par des modèles GMM, et le regroupement ILP, proposé récemment, avec lequel les classes sont représentées par des modèles i-vector.

Nous présentons également l'approche de regroupement K-moyennes (K-means), très proche de l'approche ILP.

2.3.1 Modélisation statistique du locuteur

Dans les modules les plus importants de la première composante du système de SRL d'émissions, les segments et classes de locuteurs étaient représentés par des modèles mono-gaussiens. Ce type de modélisation n'est pas réputée pour sa robustesse, il était cependant difficile de faire autrement étant donné la faible quantité de données disponibles (les segments de parole étaient alors de taille réduite). À l'issue de cette première composante, la quantité de segments de parole permettant de représenter une classe de locuteur est plus importante, de même que la longueur de ces segments. Il est donc possible de recourir à des techniques de modélisation plus élaborées pour effectuer l'étape de regroupement en locuteurs de la seconde composante du système de SRL d'émissions.

Un modèle de locuteur correspond à la représentation des caractéristiques acoustiques de la voix d'une personne. Ces modèles sont d'une importance capitale quelque soit l'application visée dans le domaine du traitement automatique de la parole, en particulier pour les applications axées locuteur (segmentation et regroupement en locuteurs, vérification du locuteur, identification du locuteur). Le modèle de locuteur idéal devrait permettre une identification incontestable de la personne qui parle. Il n'existe cependant pas, à ce jour, de méthodes suffisamment robustes [Bonastre et al., 2003a; Campbell et al., 2009]. Les méthodes de modélisation du locuteur prédominantes, fondées sur des approches statistiques et probabilistes, proviennent essentiellement du domaine de la vérification du locuteur. Dans cette partie, nous abordons les deux principales techniques employées en SRL : les modèles de mélanges gaussiens (GMM) et la modélisation i-vector.

▷ Modèles de mélanges gaussiens – GMM

L'approche GMM est devenue une référence en matière de modélisation du locuteur, et ce depuis son introduction en reconnaissance de la parole [Reynolds, 1992; Rose et Reynolds, 1990]. L'approche GMM permet une estimation plus robuste des distributions des vecteurs acoustiques, et donc, une meilleure modélisation que les modèles mono-gaussiens. Un modèle de locuteur GMM est obtenu à partir d'un ensemble de vecteurs acoustiques propres au locuteur, en réalisant une combinaison linéaire pondérée de plusieurs distributions mono-gaussiennes. La fonction de densité de distribution d'un vecteur acoustique x s'écrit :

$$p(x|\Theta) = \sum_{i=1}^{N} w_i p_i(x|\mu_i, \Sigma_i) \tag{2.5}$$

où N est le nombre de composantes gaussiennes, w_i est le vecteur de poids des composantes gaussiennes (avec $\sum_{i=1}^{N} w_i = 1$), μ_i représente les vecteurs de moyennes des composantes gaussiennes, Σ_i représente les matrices de covariance des composantes gaussiennes, $p_i(x|\mu_i, \Sigma_i)$ est la fonction de densité de la loi gaussienne du vecteur acoustique x, et Θ correspond aux paramètres du mélange gaussien (avec $\Theta = (w_i, \mu_i, \Sigma_i)$ avec $i = 1 \ldots N$).

La principale difficulté de l'approche GMM réside dans l'estimation du paramètre Θ (c'est-à-dire, w_i, μ_i et Σ_i). Déterminer les valeurs idéales de ces paramètres est une étape cruciale pour générer des modèles de locuteurs robustes. Deux méthodes sont généralement utilisées pour estimer la valeur de ces paramètres à partir de données d'apprentissage : l'algorithme espérance-maximisation (Expectation-Maximisation – EM), et l'adaptation maximum *a posteriori* (MAP) à partir d'un modèle du monde.

Algorithme EM

L'algorithme EM, proposé par [Dempster et al., 1977], est un procédé itératif visant à estimer, par maximum de vraisemblance (*Maximum Likelihood* – ML), les valeurs optimales du paramètre Θ. À chaque itération k se succèdent deux étapes permettant d'estimer un nouveau paramètre Θ_{k+1}, optimisé en fonction des données d'apprentissage et du paramètre Θ_k courant, tel que pour un vecteur acoustique x observé, $p(x|\Theta_{k+1}) \geq p(x|\Theta_k)$.

- La première étape correspond à l'évaluation de l'espérance (étape E), c'est-à-dire, calculer la probabilité *a posteriori* des composantes gaussiennes d'un vecteur acoustique observé x, pour un GMM de paramètre Θ_k.
- La deuxième étape, la maximisation (étape M), augmente le maximum de vraisemblance du paramètre Θ_k en fonction de la probabilité déterminée à l'étape E.

Les étapes E et M se répètent itérativement jusqu'à convergence vers un maximum local (la croissance de la vraisemblance étant garantie par une fonction auxiliaire). Le processus s'interrompt lorsque certaines contraintes portant généralement sur le nombre d'itérations ou sur la valeur de la vraisemblance sont satisfaites. Cette approche présente toutefois des inconvénients. L'estimation de la vraisemblance des paramètres ne convergeant que vers un maximum local, l'optimalité des paramètres

n'est pas garantie. De plus, la qualité de l'estimation des paramètres par l'algorithme EM dépend fortement de la quantité de données d'apprentissage disponible. Le nombre de composantes désirées dans le modèle GMM peut également représenter un obstacle. Plus le nombre de composantes est élevé, plus l'approximation de la distribution nécessite l'estimation d'un nombre important de paramètres. Ce dernier inconvénient peut néanmoins être contré en ne considérant, dans l'algorithme EM, que la diagonale des matrices de covariance Σ_i.

Adaptation MAP

L'adaptation MAP [Gauvain et Lee, 1994; Reynolds et al., 2000a] permet d'estimer les paramètres d'un modèle GMM dans le cas où la quantité de données d'apprentissage n'est pas suffisamment conséquente. Cette méthode tire profit des informations d'un modèle GMM connu *a priori*, dénommé *modèle du monde* (*Universal Background Model* – UBM) [Carey et Parris, 1992]. Un GMM-UBM est généralement entraîné par l'algorithme EM à partir d'une très grande quantité de données d'apprentissage correspondant à de nombreux locuteurs différents. Généralement composé de 1024 ou 2048 composantes gaussiennes, un GMM-UBM se veut capable de modéliser une distribution des caractéristiques acoustiques indépendante du locuteur. Un GMM-UBM peut être dépendant ou indépendant du genre et de la bande de fréquence, en fonction des applications visées et des données d'apprentissage utilisées.

L'approche MAP vise à adapter le paramètre $\Theta = (w_i, \mu_i, \Sigma_i)$ d'un GMM-UBM aux données du locuteur à modéliser. Ainsi, les locuteurs ne disposant pas de données en quantité suffisante ne seront pas modélisés médiocrement. L'adaptation MAP est un procédé itératif semblable à l'algorithme EM. La première étape, l'évaluation de l'espérance (étape E), se caractérise par l'utilisation du paramètre Θ_{ubm} du GMM-UBM à la place du paramètre Θ_k :

$$p(x|\Theta_{ubm}) = \sum_{n=1}^{N} p(i|x_n, \Theta_{ubm}) \tag{2.6}$$

Il a été démontré expérimentalement que les meilleurs modèles GMM sont obtenus par la simple adaptation des moyennes du GMM-UBM [Reynolds et al., 2000a], les poids et matrices de covariance du GMM-UBM restent ainsi inchangés. L'étape de maximisation (étape M) se résume alors au calcul de μ_i :

$$\mu_i = \alpha_i \frac{p(x|\Theta_{ubm})\ x_n}{p(x|\Theta_{ubm})} + (1 - \alpha_i)\mu_{i,ubm} \tag{2.7}$$

où α_i, calculé pour chaque composante, est le coefficient d'adaptation déterminant le ratio entre les anciennes et les nouvelles statistiques. Ce coefficient α_i est caractérisé par un facteur de pertinence γ prédéfini :

$$\alpha_i = \frac{p(x|\Theta_{ubm})}{p(x|\Theta_{ubm}) + \gamma} \tag{2.8}$$

L'adaptation MAP est donc caractérisée par la présence de ce facteur de pertinence γ permettant de rendre compte du degré d'adaptation des paramètres du GMM-UBM sur les données du locuteur à modéliser. Plus la valeur de ce facteur de pertinence est élevée (γ tend vers 1), plus l'estimation du vecteur de moyennes μ_i de Θ sera proche de la valeur μ_{ubm} du GMM-UBM. Au contraire, si la valeur de ce facteur de pertinence tend vers 0, alors l'estimation du paramètre μ_i de Θ sera proche de la solution donnée par la méthode EM traditionnelle. Ce facteur de pertinence permet donc d'ajuster le degré d'adaptation en fonction de la quantité de données disponibles pour modéliser les locuteurs. Le calcul des vecteurs moyens, pour chaque composante i du GMM, peut donc finalement se résumer à :

$$\mu_i = \frac{\sum_{n=1}^{N} p(i|x_n, \Theta_{ubm})\ x_n + \gamma \mu_{i,ubm}}{\sum_{n=1}^{N} p(i|x_n, \Theta_{ubm}) + \gamma} \tag{2.9}$$

À l'issue du processus d'adaptation MAP, chacune des N composantes gaussiennes du GMM est représentée par un vecteur de moyennes adaptées μ_i avec $1 \leq i \leq N$. Cet ensemble de vecteurs de moyennes peut être exprimé sous la forme d'un supervecteur $s = [\mu_1, \mu_2, \ldots, \mu_N]$ correspondant à leur concaténation.

L'adaptation MAP et les modèles GMM-UBM ont permis d'obtenir des résultats remarquables dans la plupart des applications reposant sur la modélisation du locuteur [NIST, 2002, 2004; Reynolds et al., 2000a], et la concaténation des vecteurs de moyennes sous forme de supervecteur a permis l'élaboration de techniques de décomposition en facteurs, dont l'approche de modélisation i-vector, discutée dans la partie suivante, est issue.

▷ Modélisation i-vectors

L'approche i-vector est une technique de modélisation du locuteur récente, d'abord introduite dans le domaine de la reconnaissance du locuteur [Dehak et al., 2011] puis utilisée en SRL dans le contexte du regroupement en locuteur [Rouvier et Meignier, 2012; Shum et al., 2011]. La modélisation i-vector vise à réduire la grande

quantité de données acoustiques représentant un locuteur en un vecteur discriminant de dimension réduite. Cette approche de modélisation s'inspire des méthodes de décomposition de facteurs et de réduction de dimensionnalité [Burget et al., 2007; Kenny et al., 2003; Kuhn et al., 1998; Matrouf et al., 2007], en particulier, de la méthode JFA *(Joint Factor Analysis)* [Kenny et al., 2007].

Méthode JFA

L'approche JFA repose sur l'intuition qu'un supervecteur m correspond à la somme de trois composantes indépendantes : une composante spécifique au locuteur Vy, une composante spécifique au canal Ux, et la dernière, une composante résiduelle Dz. Le modèle JFA d'une session h (session est synonyme de variabilité, elle est formée des deux composantes Ux et Dz) appartenant au locuteur s peut ainsi être exprimé par la relation suivante :

$$m_{(s,h)} = m + Ux_h + Vy_s + Dz_s \tag{2.10}$$

où m représente le supervecteur d'un GMM-UBM, à la fois indépendant du locuteur et de la session. $m_{(s,h)}$ représente le supervecteur dépendant de la session h du locuteur s. V et D désignent les sous-espaces du locuteur. V désigne la matrice de variabilité associée au locuteur *(eigenvoice matrix)*, avec y_s le facteur correspondant au locuteur de la session h. D désigne le résidu, une matrice diagonale définissant le sous-espace du locuteur *(diagonal residual)*, avec z_s le vecteur résiduel du locuteur. U désigne la matrice de variabilité associée au canal de transmission *(eigenchannel matrix)*, avec x_h un vecteur correspondant au facteur canal de la session h. U et V sont des matrices de dimensions réduites, D est incompressible, et x_h, y_s, z_s sont des vecteurs de dimension réduits suivant une distribution normale $\mathcal{N}(0, I)$.

L'algorithme permettant d'appliquer la méthode JFA consiste à estimer les composantes U_{x_h}, V_{y_s} et D_{z_s} à partir de données d'apprentissage correctement annotées en locuteur, puis à estimer les facteurs x, y et z pour une observation donnée. Le score d'un modèle JFA est déterminé par sa vraisemblance [Glembek et al., 2009] avec un modèle de référence dans lequel la composante canal est retirée ($\hat{m_{(s,h)}} = m + Vy_s + Dz_s$).

Espace de variabilité totale et i-vector

L'approche JFA permet de définir deux composantes distinctes, la composante canal, définie par la matrice U et la composante locuteur, définie par la matrice V.

[Dehak, 2009; Dehak et al., 2011; Matrouf et al., 2008] proposent une approche, motivée par l'intuition que le facteur canal de l'approche JFA contient également des informations relatives au locuteur, dans laquelle les composantes U et V sont réunies, sans aucune distinction, au sein d'un même sous-espace nommé *espace de variabilité totale*. Ce sous-espace est défini par une matrice, dite *matrice de variabilité totale*, contenant les vecteurs propres dont les valeurs propres, dans la matrice de covariance de variabilité totale, sont les plus élevées. En considérant que les variabilités du canal et du locuteur sont indépendantes, les facteurs de variabilité totale peuvent être exprimés par la fusion des composantes V et U, en ignorant le facteur résiduel D du locuteur :

$$m_{(s,h)} = m + Tw_{s,h} \tag{2.11}$$

où m désigne le supervecteur d'un GMM-UBM, à la fois indépendant du locuteur et de la session. T désigne la matrice rectangulaire de variabilité totale, de dimension réduite, réunissant les composantes canal et locuteur. Cette matrice T permet de dégager un unique facteur $w_{s,h}$, correspondant au facteur de variabilité totale normalement distribué sur $\mathcal{N}(0, I)$. L'apprentissage de la matrice T est semblable à l'apprentissage de la matrice de variabilité associée au locuteur V, en considérant toutefois les n modèles du jeu d'apprentissage comme provenant de n locuteurs distincts. Le facteur $w_{s,h}$, dénommé *i-vector*, est extrait selon l'algorithme de l'approche FA *(Factor Analysis)* en considérant la particularité de la matrice T. Le i-vector, issu d'une décomposition factorielle et d'une réduction importante de dimensionnalité d'un supervecteur s, permet d'éliminer une quantité d'information considérable et surpasse en performance les modèles de locuteurs présentement utilisés.

Normalisation des modèles

Il est commun, avec l'approche de modélisation i-vector, de réaliser le procédé de normalisation durant la phase d'évaluation, contrairement à l'approche de modélisation JFA où la normalisation des modèles est effectuée directement sur le superverteur. Cette particularité s'explique par la relative faible dimension des modèles i-vector. Dans ce manuscrit nous présentons les variantes EFR *(Eigen Factor Radial)* et SNN *(Spherical Nuisance Normalisation)* d'un algorithme de standardisation et normalisation pour les modèles i-vector. L'approche de normalisation WCCN *(Within Class Covariance Normalization)* [Dehak et al., 2011], qui est essentiellement employée lorsque la mesure de similarité entre deux modèles *i-vector* est estimée par la distance cosinus, n'est pas abordée dans ce manuscrit. Il s'agit cependant d'une

méthode de normalisation ayant largement fait ses preuves dans le domaine de la reconnaissance du locuteur.

Eigen Factor Radial (EFR) et *Spherical Nuisance Normalisation* (SNN) correspondent à deux variantes du même algorithme itératif de transformation permettant d'évaluer un ensemble de modèles i-vector $\mathcal{E}$. Cette approche de normalisation, présentée dans [Bousquet et al., 2011, 2012], requiert l'utilisation d'un ensemble $\mathcal{T}$ de modèles i-vector d'apprentissage. Lors de chaque itération, l'algorithme EFR détermine la moyenne μ_i et la matrice de covariance Σ_i des modèles i-vector de $\mathcal{T}$. Ces modèles sont alors standardisés et normalisés par division par leur norme euclidienne (Algorithme 1).

Algorithme 1 : pour les modèles i-vector d'apprentissage

for *i=1* **to** n^{bre} *d'itérations (entre 1 et 5)* **do**
  1 : Calculer la moyenne μ_i et la matrice de covariance Σ_i de l'ensemble $\mathcal{T}$
  2 : Mettre à jour les modèles i-vector d'apprentissage :
  for each w **in** $\mathcal{T}$ **do**

$$w \leftarrow \frac{\Sigma_i^{-\frac{1}{2}}(w-\mu_i)}{||\Sigma_i^{-\frac{1}{2}}(w-\mu_i)||}$$

  end
end

Chacun des modèles i-vector de l'ensemble d'évaluation $\mathcal{E}$ suit alors le même procédé de standardisation et de normalisation, en utilisant les moyennes μ_i et les matrices de covariance Σ_i obtenues durant les itérations sur les modèles de l'ensemble $\mathcal{T}$ (Algorithme 2).

Algorithme 2 : pour les modèles i-vector évalués

for each w **in** $\mathcal{E}$ **do**
  for *i=1* **to** n^{bre} *d'itérations (entre 1 et 5)* **do**

$$w \leftarrow \frac{\Sigma_i^{-\frac{1}{2}}(w-\mu_i)}{||\Sigma_i^{-\frac{1}{2}}(w-\mu_i)||}$$

  end
end

La méthode de normalisation SNN est identique en tout point à celle d'EFR, excepté qu'elle fait appel à la matrice de covariances intra-classes W au lieu de la matrice de covariance globale Σ.

2.3.2 Scores de vraisemblance

▷ Entre des modèles GMM

Les métriques les plus appréciées dans le contexte de regroupement en locuteurs, lorsque les classes sont représentées par des modèles GMM, sont l'entropie croisée (CE) et le rapport de vraisemblance croisé (CLR).

Entropie Croisée (CE)

Le calcul de l'entropie croisée (*Cross Entropy* – CE) entre deux modèles GMM peut être utilisé comme score de vraisemblance. L'entropie croisée a d'abord été utilisée pour le regroupement en locuteurs dans [Solomonoff et al., 1998]. Étant donné deux classes c_i et c_j, et leurs modèles respectifs M_i et M_j entraînés sur les données des classes c_i et c_j, l'entropie croisée se définie par :

$$CE(c_i, c_j) = \log \frac{L(c_i, M_i)}{L(c_i, M_j)} + \log \frac{L(c_j, M_j)}{L(c_j, M_i)} \tag{2.12}$$

où $L(c_i, M_j)$ correspond à la vraisemblance des données de la classe c_i par rapport au modèle M_j.

Rapport de Vraisemblance Croisé (CLR)

Reynolds et al. [1998] font appel au rapport de vraisemblance croisé (*Cross-Likelihood Ratio* – CLR) pour estimer la dissimilaritée entre deux classes c_i et c_j à partir de leurs modèles respectifs M_i et M_j :

$$CLR(c_i, c_j) = \log \frac{L(c_i, M_{UBM})}{L(c_i, M_j)} + \log \frac{L(c_j, M_{UBM})}{L(c_j, M_i)} \tag{2.13}$$

où $L(c_i, M_j)$ est la vraisemblance des données de la classe c_i par rapport au modèle M_j, et M_{UBM} est le modèle du monde employé pour l'apprentissage des modèles M_i et M_j.

La plupart des systèmes de SRL d'émissions ont recourt au rapport de vraisemblance croisé pour estimer la similarité entre les classes modélisées par des GMM, cependant, il a été constaté que l'entropie croisée procure une meilleure segmentation [Le et al., 2007] (au moins sur les corpus des campagnes ESTER). La différence

entre CLR et CE repose sur les numérateurs, qui, pour la métrique CLR correspondent aux scores de log-vraisemblance du GMM-UBM, et pour la métrique CE, aux scores de log-vraisemblance des classes.

▷ Entre des modèles i-vector

Les scores étudiés pour déterminer la proximité de deux modèles i-vectors w_i et w_j, représentant deux classes c_i et c_j, sont la distance de Mahalanobis et le score de vraisemblance PLDA (Probabilistic Linear Discriminant Analysis). La similarité cosinus [Dehak et al., 2011], habituellement employée avec la méthode de normalisation WCCN, n'est pas abordée dans ce manuscrit.

Score Mahanalobis

Une fois la normalisation des modèles i-vector de test effectuée, [Bousquet et al., 2011] propose de recourir au score Mahalanobis afin de déterminer si deux modèles i-vector w_i et w_j correspondent à un même locuteur, ou non :

$$s_{maha} = -\frac{1}{2}(w_i - w_j)^t W^{-1}(w_i - w_j) \tag{2.14}$$

où W correspond à la matrice de covariance intra-classe calculée à partir des modèles i-vector de l'ensemble d'apprentissage (après application de la normalisation EFR). W^t est fréquemment appelée *matrice de Mahalanobis*. Cette matrice W correspond à la moyenne des matrices de covariance des locuteurs. Elle est calculée à partir de n modèles i-vector de l'ensemble d'apprentissage selon l'équation suivante :

$$W = \frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n_s}(w_i^s - \bar{w}_s)\,(w_i^s - \bar{w}_s)^t \tag{2.15}$$

où n_s correspond au nombre de sessions du locuteur s, S correspond au nombre total de locuteurs, w_i^s est le modèle i-vector du corpus d'apprentissage du locuteur s et de la session i, et $\bar{w}_s$ correspond à la moyenne des modèles i-vector du locuteur s.

Finalement, le score Mahalanobis peut être lu comme une distance entre deux vecteurs par la relation :

$$d_{maha} = -\frac{1}{2}||w_i - w_j||^2_{W^{-1}} \tag{2.16}$$

Score PLDA

En vérification du locuteur, l'analyse discriminante linéaire probabiliste (*Probabilistic Linear Discriminant Analysis* – PLDA) a été adaptée du domaine de l'imagerie [Prince et Elder, 2007] de manière à modéliser la distribution des modèles i-vector [Kenny, 2010]. Cette approche, qui peut être considérée comme un cas particulier mono-gaussien de l'approche JFA [Kenny et al., 2008], est désormais largement employée en reconnaissance du locuteur [Jiang et al., 2012; Matejka et al., 2011]. L'apprentissage d'un modèle PLDA nécessite un corpus d'apprentissage constitué de plusieurs enregistrements (ou *sessions*) d'un grand nombre de locuteurs. Le modèle i-vector $w_{s,h}$, qui modélise la session h du locuteur s, est exprimé par la décomposition de facteurs suivante :

$$w_{s,h} = \mu + Vy_s + Uz_{s,h} + \epsilon_{s,h} \tag{2.17}$$

où μ représente la moyenne globale, à la fois indépendante du locuteur et de la session, des modèles i-vector du corpus d'apprentissage. Les espaces de variabilité du locuteur et de la session sont respectivement représentés par les colonnes de la matrice V et de la matrice U. $\epsilon_{s,h}$ représente la variabilité résiduelle modélisée par une matrice de covariance pleine. Le terme y_s correspond au facteur locuteur *(eigenvoice factor)*, et le terme $z_{s,h}$ représente le facteur canal de la session h pour le locuteur s. Ces deux facteurs sont supposés indépendants et suivent une distribution normale sur $\mathcal{N}(0, I)$. Les paramètres $\{\mu, V, U, \epsilon\}$ du modèle PLDA sont estimés itérativement par l'algorithme EM.

À la manière de la méthode de rapport de vraisemblance généralisé, deux hypothèses sont testées :

- $\mathcal{H}_0$: les deux modèles i-vector w_i et w_j sont produits par un même locuteur, auquel cas ces modèles partagent le même facteur locuteur y_s.
- $\mathcal{H}_1$: les deux modèles i-vector w_i et w_j sont produits par deux locuteurs différents, auquel cas le facteur locuteur y_s de ces deux modèles est différent.

Le score de vraisemblance PLDA entre deux modèles i-vector normalisés w_i et w_j, calculé pour les classes c_i et c_j, est proportionnel au logarithme de la probabilité que w_i et w_j appartiennent au même locuteur :

$$S_{PLDA} = \log \frac{p(w_i, w_j|M_0)}{p(w_i, w_j|M_1)} \tag{2.18}$$

où, sous le modèle M_0 les modèles i-vector w_i et w_j ont été produits par le même

locuteur (hypothèse $\mathcal{H}_0$), et sous le modèle M_1 les modèles i-vector w_i et w_j sont produits par deux locuteurs différents (hypothèse $\mathcal{H}_1$).

2.3.3 Regroupements hiérarchiques

Le regroupement hiérarchique est une méthode de classification qui se décline en deux stratégies : on distingue l'approche agglomérative – ou ascendante – *(bottom-up)*, et l'approche descendante *(top-down)*. Le principe de ces deux stratégies de regroupement est illustré en figure 2.6.

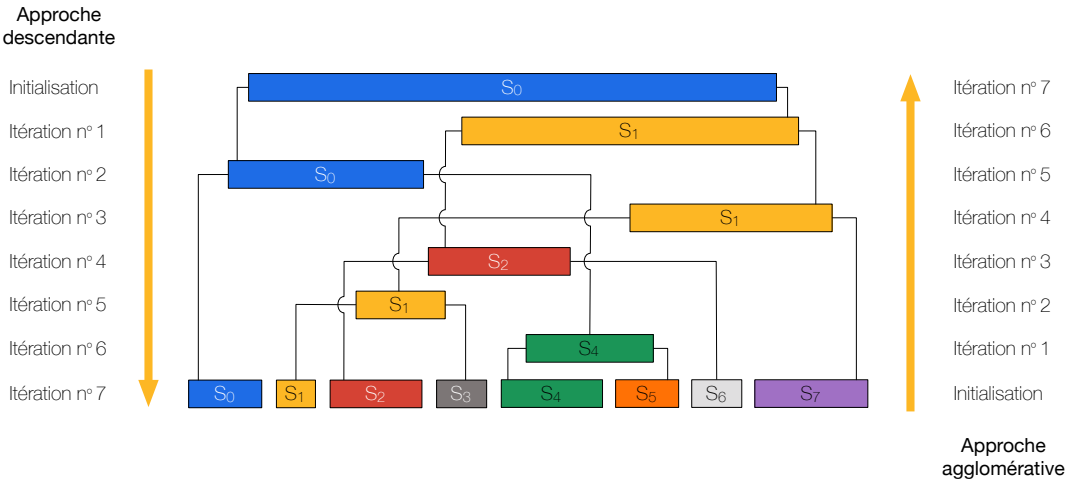

Figure *2.6 – Exemple de regroupement hiérarchique présentant les approches agglomérative (à droite) et descendante (à gauche).*

L'approche agglomérative démarre avec un ensemble initial de classes distinctes qui vont être itérativement regroupées en fonction de leur vraisemblance. Dans le contexte de la SRL, les classes sont représentées par des modèles de locuteurs, et les similarités sont établies en fonction de leurs caractéristiques acoustiques. L'approche descendante suit le raisonnement inverse, elle démarre avec une unique classe qui va successivement être divisée en sous-classes en fonction de leur non-vraisemblance (établies *a posteriori*). L'approche agglomérative est largement préférée à l'approche descendante dans le domaine de la SRL, en particulier pour la simplicité avec la laquelle une segmentation peut servir d'entrée au processus.

▷ Approche descendante

L'approche descendante, rarement utilisée en SRL, a cependant été implémentée avec succès dans le cadre des évaluations NIST pour la reconnaissance automatique

de la parole [Johnson et Woodland, 1998]. La méthode proposée avait pour objectif de regrouper les segments aux caractéristiques acoustiques proches, de manière à optimiser les performances de l'adaptation en locuteurs par régression linéaire de vraisemblance maximum (*Maximum Likelihood Linear Regression* – MLLR) [Leggetter et Woodland, 1995]. La méthode présentée dans [Johnson et Woodland, 1998] consiste à diviser itérativement les données d'une classe en quatre sous-classes, tant que la quantité de données le permet, puis à fusionner les classes les plus proches. Cette méthode a été adaptée aux besoins de la SRL [Anguera et Hernando, 2004; Johnson, 1999; Meignier et al., 2001; Tranter et Reynolds, 2004; Tranter et al., 2004]. Pour être efficace, l'adaptation MLLR impose une durée minimum de parole par classe de locuteur, quitte à en diminuer la pureté. Cette contrainte entre en contradiction avec l'objectif principal de la SRL, qui est la production de classes de locuteurs les plus pures possible, même pour les locuteurs ayant un temps de parole très court.

▷ Approche agglomérative

Le regroupement agglomératif hiérarchique est probablement la méthode de classification automatique la plus employée dans le domaine de la SRL [Chen et Gopalakrishnan, 1998; Gish et al., 1991; Reynolds et al., 1998; Siegler et al., 1997; Siu et al., 1992; Solomonoff et al., 1998].

Ce procédé de classification itératif démarre avec un ensemble de n classes initiales, une pour chacune des classes de locuteurs (ensemble de segments de parole) déterminées à l'issue de la première composante du système de SRL d'émissions. Lors de chaque itération, l'algorithme va successivement estimer les mesures de vraisemblance entre chaque paire de classes, puis regrouper les deux classes les plus proches. L'algorithme s'interrompt lorsque toutes les classes ont été regroupées en une seule et unique classe, ou lorsqu'un critère d'arrêt est rencontré.

Les éléments importants à prendre en compte dans la classification hiérarchique agglomérative sont donc, d'une part, le score de vraisemblance permettant d'estimer la proximité des classes, et d'autre part, le critère d'arrêt à partir duquel le processus de regroupement s'interrompt. Ces deux facteurs ont une influence importante sur la qualité de la classification. Un score de vraisemblance inadapté provoquera de mauvais regroupements en locuteurs. Le critère d'arrêt est quant à lui déterminant pour ne pas aller trop loin dans le regroupement. Sans ce critère d'arrêt, l'approche agglomérative finira par regrouper toutes les classes de l'ensemble initial en une seule et unique classe.

Estimation de la vraisemblance

À chaque itération de la classification hiérarchique agglomérative, les deux classes les plus proches, c'est-à-dire celles qui maximisent le score de vraisemblance, sont regroupées. Il est donc nécessaire, à chaque itération, de déterminer les scores de vraisemblance entre la nouvelle classe, issue du regroupement de l'itération précédente, et les classes restantes. Deux stratégies ont été proposées pour estimer les scores de vraisemblance entre des classes issues de regroupement (des classes constituées de plusieurs éléments). La première, la plus commune dans le domaine de la reconnaissance du locuteur, consiste à considérer la nouvelle classe comme la représentation d'un unique et large segment, en concaténant tous les segments de cette classe et constituant un nouveau modèle de locuteur pour la représenter. Il s'agit d'une stratégie coûteuse étant donné qu'elle implique la lecture des paramètres acoustiques (MFCC) pour constituer le modèle de locuteur. La deuxième stratégie, plus commune en classification de données, tient compte de la distance entre les segments qui constituent les classes. À ce titre, plusieurs critères de liaison peuvent être envisagés [Solomonoff et al., 1998] mais seuls les critères de liaison minimum et maximum fonctionnent quel que soit le score de vraisemblance utilisé, qu'il s'agisse, ou non, de distances (la propriété d'inégalité triangulaire n'est pas impérative) :

- Liaison minimum : le score de vraisemblance entre deux classes c_i et c_j correspond au score le plus faible entre les segments de la classe c_i de ceux de la classe c_j.
- Liaison maximum : le score de vraisemblance entre deux classes c_i et c_j correspond au score le plus élevé entre les segments de la classe c_i de ceux de la classe c_j.

Critère d'arrêt

Chacune des itérations de l'algorithme de regroupement diminue le nombre de classes en les agglomérant les unes aux autres. L'algorithme se termine naturellement lorsque toutes les classes ont été agglomérées. Cependant, l'étape de regroupement en locuteurs a pour objectif d'agglomérer les segments d'un même locuteur au sein d'une même classe, il est donc nécessaire d'interrompre le processus de regroupement avant qu'il ne fusionne deux classes correspondant à des locuteurs différents. Pour ce faire, on définit un « critère d'arrêt » permettant de terminer prématurément l'algorithme de regroupement. Ce critère d'arrêt correspond à une contrainte forte portant généralement sur la valeur du score de vraisemblance ou, lorsque le nombre de locuteurs est connu *a priori*, sur le nombre de classes.

En SRL, sur les émissions journalistiques, le nombre de locuteurs présents dans l'enregistrement audio est inconnu. Le critère d'arrêt utilisé correspond alors à la valeur du score de vraisemblance à partir de laquelle les classes ne devraient plus être regroupées, car trop éloignées les unes des autres. Ce *seuil*, généralement déterminé empiriquement, dépend du score de vraisemblance choisi et des conditions acoustiques d'enregistrement.

Discussion

Le regroupement agglomératif hiérarchique est une technique de regroupement en locuteurs très répandue. Cette approche, utilisée conjointement à une modélisation du locuteur par modèles de mélanges gaussiens (GMM), a permis d'obtenir de très bons résultats dans les principales campagnes d'évaluation en SRL sur les émissions journalistiques d'information, telles que RT-04F [NIST, 2004], REPERE [Galibert et Kahn, 2013], Albayzin [Zelenák et al., 2012] et ESTER 2 [Galliano et al., 2009]. Le regroupement agglomératif hiérarchique souffre cependant de deux inconvénients majeurs.

Premièrement, la complexité algorithmique. À chaque itération, un nouveau modèle de locuteur doit être calculé pour représenter la nouvelle classe issue du regroupement. Plus le processus de regroupement est avancé, plus les modèles de locuteurs à calculer sont conséquents, du fait de l'agglomération des segments, et donc, de la quantité de données disponible pour la modélisation. Les scores de vraisemblance entre le nouveau modèle de locuteur ainsi obtenu, et ceux qui représentent les classes restantes, doivent être déterminés. Certains compromis peuvent être faits pour réduire cette complexité. Par exemple, les modèles de locuteur de type GMM peuvent être obtenus via une seule itération de l'algorithme MAP. Le nouveau modèle GMM résulte alors de la fusion des accumulateurs statistiques des modèles GMM correspondants aux classes regroupées. Pour accélérer le calcul des scores de vraisemblance, il est également possible de ne considérer que les n premières composantes gaussiennes des modèles GMM, auquel cas, seulement n regroupements sont effectués. Malgré ces deux compromis, l'approche de regroupement hiérarchique ascendante reste lente, en particulier lorsque le nombre de classes impliquées est élevé, et que le seuil correspondant au critère d'arrêt est faible.

Deuxièmement, la propagation des erreurs. Le regroupement hiérarchique ne permet pas une vision globale du problème de regroupement, il ne permet d'explorer qu'une seule solution parmi l'ensemble des possibles. Bien que les classes à regrouper lors de chaque itération soient déterminées en fonction de la vraisemblance maximale parmi tous les couples de classes, il ne s'agit que de regroupements optimaux

locaux. Les regroupements, une fois réalisés, ne sont jamais remis en question. Par conséquent, un regroupement erroné sera propagé jusqu'à la rencontre du critère d'arrêt, pouvant mener à d'autres regroupements incorrects, et finalement dégrader la classification. Des solutions existent pour pallier à ce problème de propagation des erreurs, les systèmes intégrés où segmentation et classification sont itérativement enchaînées de manière à remettre en cause les décisions prises [Meignier et al., 2001, 2006]. La convergence vers une classification optimale n'est cependant pas garantie.

▷ Configuration pour le regroupement hiérarchique

Avec cette seconde composante de l'architecture visant à optimiser les segmentations produites pour satisfaire les critères de la tâche de SRL, les classes de locuteurs sont représentées par des modèles GMM et l'approche de regroupement hiérarchique employée est l'approche agglomérative.

La particularité avec ce nouveau regroupement agglomératif hiérarchique est la normalisation des paramètres acoustiques pour retirer la contribution du canal, au moyen des méthodes *features warping* et MVN. Lors du regroupement hiérarchique BIC, effectué durant la première composante de l'architecture (*cf.* partie 2.2.3), les paramètres acoustiques n'étaient pas normalisés, car la contribution du canal était un indice clé pour différencier les locuteurs.

Les modèles GMM utilisés pour représenter les classes de locuteurs ont été obtenus par l'intermédiaire de l'adaptation MAP d'un modèle du monde. Ce GMM-UBM à 512 composantes gaussiennes correspond à la concaténation des quatre modèles GMM utilisés lors de l'étape de détection du genre et de la bande de fréquence, dans la première composante de l'architecture. Le score de vraisemblance utilisé pour estimer la similarité entre les classes correspond à l'entropie croisée (CE).

2.3.4 Regroupements combinatoires

Le problème de complexité algorithmique évoqué dans la partie précédente est une réalité avec laquelle des compromis peuvent être faits. Le problème de propagation des erreurs au fil des itérations du regroupement agglomératif hiérarchique peut être géré par un processus itératif alternant segmentation et regroupement (systèmes intégrés), cependant, le processus reste coûteux en temps et incertain en termes de classification. Des approches de regroupements *combinatoires* ont été récemment proposées pour s'affranchir de cet inconvénient lié à la propagation des erreurs sans

avoir recours aux approches intégrées. Dans cette partie, le regroupement en locuteurs est considéré comme un problème de partitionnement, où n segments (n classes initiales) doivent être répartis dans c classes à découvrir, c correspondant idéalement au nombre de locuteurs présents dans l'enregistrement. Le problème se résume donc à déterminer la meilleure partition parmi l'ensemble des possibles.

L'inconvénient majeur avec cette vision du regroupement en locuteurs est qu'il s'agit d'un problème combinatoire de type *NP*-complet, dans certains cas impossibles à résoudre [Cook, 2006]. De ce fait, la solution optimale ne peut être qu'approchée par des algorithmes d'approximation ou des heuristiques de recherche efficaces. Le regroupement en locuteurs, ainsi formulé, permet néanmoins de considérer le problème dans sa globalité : on recherche une solution optimale pour l'ensemble du problème, contrairement au regroupement hiérarchique qui recherche, itérativement, une solution optimale pour un sous-ensemble du problème. Jusqu'à présent, deux principales méthodes de regroupement combinatoire ont été proposées : la première exploite l'algorithme k-moyennes *(*k*-means)*, et la seconde propose une généralisation exprimée en Programmation Linéaire en Nombres Entiers (*Integer Linear Programming* – ILP).

▷ Regroupement k-moyennes

Le regroupement k-moyennes *(*k*-means)* a été appliqué dans le cadre de la SRL sur des enregistrements téléphoniques [Shum et al., 2011]. L'algorithme k-moyennes vise à diviser un nombre n d'observations (dans le contexte de la SRL, des segments) en k partitions (classes), le nombre k devant être déterminé *a priori*. Étant donné un ensemble de segments $\mathbf{S} = \{s_1, \ldots, s_n\}$, où chaque segment est représenté par un vecteur de dimension d, l'algorithme k-moyennes cherche à partitionner les n segments en k partitions, avec $\mathbf{P} = \{P_1, \ldots, P_k\}$ $(k \leq n)$, tout en minimisant la dispersion intra classe :

$$\arg\min_{\mathrm{P}} \sum_{i=1}^{k} \sum_{\mathrm{s}_j \in P_i} \|\mathbf{s}_j - \boldsymbol{\mu}_i\|^2 \tag{2.19}$$

où μ_i est la moyenne vectorielle des segments dans la partition P_i.

L'algorithme est itératif et s'interrompt lorsque la convergence est atteinte. En raison de la nature complexe du problème de partitionnement, des critères d'arrêt peuvent être définis (en général, sur le nombre d'itérations). L'inconvénient principal du regroupement k-moyennes réside dans l'initialisation du paramètre k. En effet, le nombre de partitions doit être connu *a priori*, ce qui ne pose pas problème dans

le cadre d'enregistrements téléphoniques entre deux locuteurs ($k = 2$). Lorsque le nombre de partitions n'est pas connu à l'avance, comme c'est le cas dans la SRL sur des émissions journalistiques d'information – le nombre de locuteurs est inconnu –, il est nécessaire d'estimer préalablement la valeur de k. Une autre approche consiste à tester différentes valeurs de k et sélectionner celle qui minimise la dispersion.

Dans l'approche présentée par [Shum et al., 2011], les segments sont modélisés par des modèles i-vector de dimension 400, et la mesure de distance utilisée pour minimiser la variabilité intra classe est la similarité cosinus. Dans [Shum et al., 2012], une estimation du nombre de partitions est proposée par l'intermédiaire d'un regroupement spectral [Ning et al., 2006], réalisé en amont, de manière à généraliser le regroupement k-moyennes quand k est *a priori* inconnu.

▷ Regroupement ILP

Le regroupement par Programmation Linéaire en Nombres Entiers (ILP), proposé par [Rouvier et Meignier, 2012], se caractérise par une fonction objective à minimiser et par la manipulation de variables binaires. Cette approche repose sur l'hypothèse qu'un segment s appartient à une classe k si le score de vraisemblance entre le centre de la classe k et le segment s est inférieur à une certaine valeur. Dans cette approche, le centre d'une classe correspond nécessairement à l'un des segments de l'ensemble $\mathcal{S} = \{s_1, \ldots, s_n\}$, il peut donc y avoir autant de centres qu'il y a de segments. L'objectif est double : d'une part, minimiser le nombre C de classes centrales, c'est-à-dire, le nombre de segments centre de classe, et d'autre part, minimiser la dispersion des segments au sein des classes. Le nombre $C \in \{1, \ldots, N\}$ est à déterminer automatiquement. Ce problème de regroupement est exprimé par une fonction objective, présentée en équation 2.20a. Les variables manipulées sont binaires : $x_{k,k}$ indique que le segment k est un centre, et $x_{k,j}$ indique le segment j appartient à la classe de centre k. La fonction objective est composée de deux parties, chacune faisant référence à l'un des objectifs fixés : $\sum_{k=1}^{N} x_{k,k}$ détermine le nombre de classes centrales C, et $\sum_{k=1}^{N}\sum_{j=1}^{N} s(k,j)x_{k,j}$ calcule la somme des scores entre le segment centroïde k et les segments appartenant à la classe de centre k. Le problème de regroupement consiste donc à minimiser la somme de ces deux fonctions : d'une part, le nombre de centres (classes), d'autre part, la dispersion intra classe.

$$\textit{Minimiser} : \qquad \sum_{k=1}^{N} x_{k,k} + \frac{1}{(S+1)} \sum_{k=1}^{N}\sum_{j=1}^{N} s(k,j)x_{k,j} \tag{2.20a}$$

$$Contraintes : \quad x_{k,j} \in \{0,1\} \qquad k \in C, j \in C \tag{2.20b}$$

$$\sum_{k=1}^{N} x_{k,j} = 1 \qquad j \in C \tag{2.20c}$$

$$x_{k,j} - x_{k,k} \leq 0 \qquad k \in C, j \in C \tag{2.20d}$$

$$s(k,j)x_{k,j} < \delta \qquad k \in C, j \in C \tag{2.20e}$$

Dans l'équation 2.20a, $s(k,j)$ correspond au score de vraisemblance entre le segment centroïde k et le segment j, et $1/(S+1)$ est un facteur de normalisation permettant de pondérer les deux parties de l'équation. S correspond à somme de tous les scores inférieurs au seuil δ déterminé expérimentalement. La contrainte 2.20b précise que les variables manipulées sont binaires : $x_{k,k}$ est égale à 1 si le segment k est centre d'une classe, et $x_{k,j}$ est égale à 1 quand le segment j est associé à la classe de centre k. La contrainte 2.20c vérifie qu'un segment j n'est associé qu'à une unique classe de centre k. La contrainte 2.20d vérifie que le segment k est sélectionné (*i.e.* $x_{k,k} = 1$) si un segment j est assigné à la classe de centre k. La contrainte 2.20e spécifie qu'un segment j peut être associé à une classe de centre k seulement si le score entre k et j est inférieur à un score δ déterminé expérimentalement.

Dans l'approche proposée par [Rouvier et Meignier, 2012] les segments sont également modélisés par des modèles i-vector, et l'appartenance de deux segments à un même locuteur est estimée par une distance de Mahalanobis. Ce problème, une fois exprimé en fonction de la segmentation fournie à l'issue de la composante de l'architecture visant à optimiser les segmentations pour la transcription automatique, est résolu par l'algorithme *Branch & Bound* du programme d'optimisation linéaire *GNU Linear Programming Toolkit*[3].

▷ Configuration pour le regroupement ILP

Avec cette approche alternative au regroupement agglomératif hiérarchique, le regroupement est exprimé sous la forme d'un problème de Programmation Linéaire en Nombres Entiers. La « recette » suivie est identique à celle proposée par [Rouvier et Meignier, 2012]. Un modèle i-vector de dimension 60, normalisé par EFR, est extrait pour chacune des classes de locuteurs à partir d'un GMM-UBM constitué de 1024 composantes gaussiennes. Ce modèle du monde a été entraîné sur les données ESTER 1. Les modèles i-vector et le GMM-UBM sont extraits à partir de

3. http://www.gnu.org/software/glpk/

la suite d'outils pour la reconnaissance du locuteur *Alize* [Bonastre et al., 2008]. La paramétrisation acoustique consiste en 19 paramètres MFCC normalisés par MVN, l'énergie, et leurs coefficients différentiels Δ et $\Delta\Delta$ respectifs.

2.3.5 Bilan

La seconde composante de notre architecture pour la SRL d'émissions consiste donc en une étape de regroupement en locuteurs pour laquelle la contribution du canal est minimisée. L'ensemble de classes initiales est déterminé par la segmentation fournie à l'issue de la première composante, qui propose des segmentations adaptées aux besoins de la tâche de transcription automatique de la parole. L'approche de regroupement agglomératif hiérarchique, associée à la modélisation GMM, reste une référence en matière de classification en locuteurs, comme en témoignent les résultats obtenus durant les récentes campagnes d'évaluation sur les émissions journalistiques d'information, telles que REPERE [Galibert et Kahn, 2013], Albayzin [Zelenák et al., 2012]. L'approche de regroupement ILP et la technique de modélisation i-vector, toutes deux beaucoup plus récentes, s'avèrent être une alternative sérieuse au regroupement hiérarchique et à la modélisation GMM. L'approche de regroupement ILP, qui a été étudiée dans différentes situations [Bredin et Poignant, 2013; Bredin et al., 2014], a été appliquée avec succès lors de la campagne REPERE, permettant d'égaler ou surpasser les résultats obtenus en SRL par l'approche hiérarchique.

2.4. Évaluation en SRL d'émissions : le DER

La métrique d'évaluation communément utilisée pour apprécier la qualité de la tâche de Segmentation et Regroupement en Locuteurs est le *Diarization Error Rate (DER)*. Le DER, défini et introduit par le *NIST* au cours des campagnes d'évaluations *Rich Transcription* [NIST, 2006], correspond à un taux d'erreur mesurant la proportion de temps de parole qui n'est pas attribuée au bon locuteur. Les segmentations produites par un système de SRL consistent en un ensemble de segments, chacun étant caractérisé par son temps de début et de fin dans l'enregistrement audio, ainsi que par une étiquette désignant le locuteur auquel il se réfère. La tâche de SRL peut être évaluée en considérant la meilleure correspondance entre les segmentations de référence, et celles fournies par le système. Cette correspondance est établie de manière à maximiser l'appariement, sur l'ensemble des locuteurs de référence, du temps de parole attribué à la fois aux locuteurs des segmentations de référence, et à leur correspondance dans les hypothèses produites par le système. Le DER est calculé sur

l'intégralité des enregistrements audio, en incluant les zones de parole superposées où plusieurs locuteurs s'expriment en même temps dans la segmentation de référence. Le DER s'exprime ainsi par la relation suivante :

$$DER = \sum_{s=1}^{\mathbf{S}} dur(s) \cdot \frac{\max(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s)}{N_{ref}} \tag{2.21}$$

où s représente un segment issu de l'ensemble des segments $\mathbf{S}$, $dur(s)$ correspond à la durée du segment s, $N_{ref}(s)$ représente le nombre de locuteurs associé au segment s dans la segmentation de référence, $N_{hyp}(s)$ représente le nombre de locuteurs associé au segment s dans la segmentation fournie par le système, et $N_{correct}(s)$ est le nombre de locuteurs associés au segment s pour lesquels une correspondance entre la segmentation de référence et celle fournie par le système a été établie. L'ensemble des segments $\mathbf{S}$ correspond à l'intersection, au sens mathématique, des segmentations de référence et d'hypothèse. Ce découpage est réalisé en fonction des temps de début et fin des segments, en considérant à la fois la segmentation de référence et celle fournie par le système. De cette manière, l'ensemble de segments $\mathbf{S}$ pris en compte par la formule est invariable en nombre et en durée (*cf.* figure 2.7).

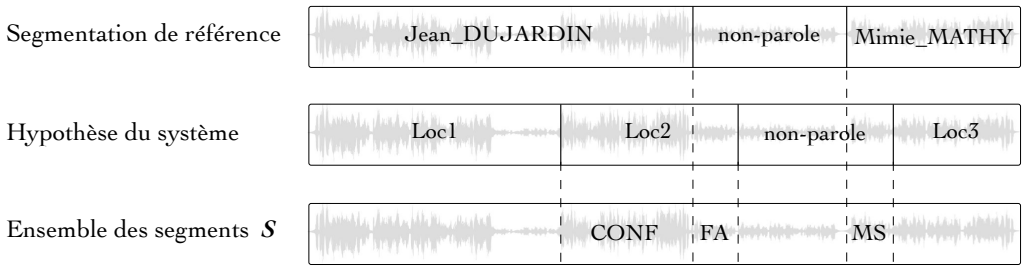

Figure *2.7 – Représentation de l'ensemble des segments* $\mathbf{S}$*, ainsi que des erreurs prises en compte dans le calcul du DER.*

Le DER, exprimé dans l'équation 2.21, peut être décomposé en trois types d'erreurs distinctes :

1. Les erreurs de type FA *(False Alarm)*, qui correspondent aux intervalles de temps de parole attribués à tort à un locuteur. Il s'agit du cas où le système détecte de la parole alors qu'il n'y en a pas. Ces erreurs peuvent être déterminées grâce à la formulation suivante :

$$FA = \sum_{s=1}^{\mathbf{S}} dur(s) \cdot \frac{N_{hyp}(s) - N_{ref}(s)}{N_{ref}} \qquad \forall N_{hyp}(s) - N_{ref}(s) > 0 \tag{2.22}$$

2. Les erreurs de type MS *(Miss)*, qui correspondent à des intervalles de temps de parole attribués à aucun locuteur. Il s'agit du cas contraire aux erreurs de type FA, où le système ne détecte pas de parole alors qu'il y en a. Ces erreurs peuvent être exprimées par la relation :

$$MS = \sum_{s=1}^{S} dur(s) \cdot \frac{N_{ref}(s) - N_{hyp}(s)}{N_{ref}} \qquad \forall N_{ref}(s) - N_{hyp}(s) > 0 \quad (2.23)$$

3. Les erreurs de type CONF *(Confusion)*, qui correspondent aux intervalles de temps de parole attribués à de mauvais locuteurs, en considérant la correspondance optimale entre les segmentations de référence et celles fournies par le système. Ces erreurs se calculent de la manière suivante :

$$CONF = \sum_{s=1}^{S} dur(s) \cdot \frac{\min(N_{hyp}(s), N_{ref}(s)) - N_{correct}(s)}{N_{ref}} \quad (2.24)$$

Les erreurs sont généralement comptées en MS ou FA, selon que les locuteurs mal affectés proviennent de la segmentation de référence ou de l'hypothèse fournie par le système. Si plusieurs locuteurs apparaissent à la fois dans la référence et dans l'hypothèse, alors l'erreur obtenue est de type CONF. Étant donné les différents types d'erreurs, l'équation 2.21 peut être reformulée de la manière suivante :

$$DER = FA + MS + CONF \quad (2.25)$$

Le dernier point à aborder sur cette métrique d'évaluation concerne les imprécisions de segmentation dans les données de référence. Il est très difficile de déterminer avec précision quand commence et quand s'arrête un tour de parole, en particulier dans les environnements bruités, ou lorsque plusieurs personnes parlent en même temps (parole superposée). Afin de prendre en compte les imprécisions de segmentation liées à l'annotation manuelle des données de références, ou à un alignement forcé, le NIST propose de réaliser l'évaluation en faisant abstraction d'un intervalle de 250 ms autour de la frontière des segments. Suivant les corpus, l'impact est non-négligeable : +1% sur ESTER, + 3% sur ETAPE, +2% en moyenne sur REPERE.

Les DER présentés dans ce manuscrit ont tous été calculés avec l'outil développé par le Laboratoire National de métrologie et d'Essais[4] (LNE) dans le cadre des campagnes d'évaluation ETAPE [Gravier et al., 2012] et REPERE [Galibert et Kahn, 2013]. Cet outil utilise l'algorithme Hongrois pour rechercher la meilleure correspondance entre les segmentations de référence et les hypothèses fournie par le système

4. http://www.lne.fr

[Galibert, 2013]. Contrairement à l'outil d'évaluation proposé par le NIST, qui met en œuvre une heuristique pour approcher la solution optimale, l'algorithme Hongrois est efficace quelque soit la quantité de parole superposée à traiter, et est adapté aux conditions d'évaluation par collections.

2.5. Bilan général sur la SRL d'émissions

L'architecture et les approches présentées lors de cet état de l'art en SRL d'émissions permettent de produire des segmentations en locuteurs généralement précises. Les récentes avancées réalisées dans le domaine sont fortement corrélées à l'évolution des techniques de modélisation en locuteurs, en reconnaissance du locuteur. Le tableau 2.1 présente, à titre informatif, les progrès clés réalisés sur les approches de modélisation en vérification du locuteur [Hagai Aronowitz, 2014].

Algorithme	Année	EER (%)
GMM	1995	> 10
GMM-UBM + score-norm [Reynolds et al., 2000b]	2000	6,2
GMM-supervectors	2004	6,2
NAP / WCCN / Eigen-channels	2005	3,6
JFA [Kenny et al., 2007]	2006	1.4
i-vectors + PLDA [Garcia-Romero et Espy-Wilson, 2011]	2011	1.0

Table *2.1 – Progrès de l'état de l'art en reconnaissance du locuteur sur les données téléphoniques NIST-SRE-10'. Les résultats sont présentés en termes de taux d'erreur EER (Equal Error Rate).*

Les résultats sont présentés en taux d'erreur EER *(Equal Error Rate)*, et les données sur lesquelles ont été calculés ces scores correspondent aux données téléphoniques de la campagne d'évaluation NIST-SRE-10' [NIST, 2010]. Concernant les approches de regroupement, on retiendra surtout l'introduction des regroupements combinatoires K-moyenne et ILP, qui coïncide avec l'introduction de l'approche de modélisation i-vector. Nous présentons à cet effet les résultats obtenus par différentes approches de regroupement dans le tableau 2.2. Ces données ont été recueillies par Sylvain Meignier dans le cadre de son HDR. L'approche de regroupement ILP, associée à une modélisation en locuteurs *i-vector* et une évaluation PLDA permet d'atteindre des résultats en termes de DER significativement inférieurs à ceux des autres approches sur différents corpus d'enregistrements de type *journaux d'informations* radio et télévisuels.

La particularité de la SRL d'émissions est de traiter séparément les différents enregistrements qui composent un corpus de données. Les segmentations produites sont donc annotées en fonction de l'enregistrement traité. Cette architecture montre

Système	E-HMM	BIC	CE/GMM	ILP/PLDA
Seuil	-	$\lambda = 6$	$\delta = 1{,}6$	$\delta = 20$
ESTER 1	NC.	16,86% (16,60%)	10,05% (7,37%)	8,87% (7,71%)
ESTER 2	NC.	13,25% (13,25%)	12,91% (9,09%)	8,87% (7,71%)
ETAPE	NC.	23,55% (23,26%)	26,31% (25,50%)	19,55% (17,70%)
REPERE 1	NC.	15,09% (14,77%)	26,31% (25,50%)	12,42% (12,12%)
RT'03 S	15,19%	9,62% (9,62%)	26,31% (25,50%)	5,23% (5,23%)
Moyenne	NC.	17,28%	15,95%	**11,51%**

Table *2.2 – Comparaison des systèmes pour le seuil donnant le meilleur résultat pour l'évaluation des 90 enregistrements (ESTER 1 & 2, ETAPE, REPERE 1, RT'03 S). (x%) : les meilleurs DER pour chaque corpus. Ces résultats proviennent du manuscrit de l'HDR de Sylvain Meignier.*

ses limites dès lors que l'on souhaite confronter les segmentations d'enregistrements différents, de manière à détecter les locuteurs qui interviendraient dans des enregistrements différents.

CHAPITRE 3

État de l'art en SRL de collections

Le chapitre 2 fait état des principales approches utilisées pour segmenter, regrouper et modéliser les locuteurs présents dans les enregistrements d'émissions journalistiques d'information. Le chapitre 3 s'intéresse aux approches développées afin de généraliser le procédé de SRL au traitement des collections d'émissions. Une *collection* est caractérisée par un ensemble d'enregistrements dans lesquels certains locuteurs sont dits « récurrents ». Un locuteur est récurrent s'il intervient dans plusieurs enregistrements de la collection, régulièrement, ou sporadiquement. La SRL de collections vise à regrouper sous une même classe, et donc sous une même étiquette, les segments correspondant à un même locuteur, et ce, quel que soit l'enregistrement de la collection dans lequel il intervient. C'est en ce sens qu'il s'agit d'une généralisation de la tâche de SRL, car si les techniques restent les mêmes, l'échelle change. Travailler sur un ensemble d'enregistrement n'est pas un concept nouveau en SRL, mais reste relativement récent. La notion d'appariement en locuteurs d'une collection a été proposée en 2002 dans le cadre de la thèse de Sylvain Meignier [Meignier, 2002; Meignier et al., 2002]. C'est à partir des années 2010 que les travaux sur la SRL de collections se sont multipliés : différentes approches se sont succédé, et les dénominations se sont diversifiées. [Bourlard et al., 2013; Ferràs et Bourlard, 2012; Ghaemmaghami et al., 2013; Leeuwen, 2010] définissent et utilisent la notion d'appariement en locuteurs *(speaker linking)*, et considèrent ce procédé comme une tâche indépendante de la SRL, effectuée *a posteriori* à partir des segmentations produites par un système de SRL d'émissions. [Tran et al., 2011; Yang et al., 2011] estiment quant à eux que cette généralisation s'intègre directement dans le procédé de SRL par l'ajout d'une étape de regroupement en locuteurs supplémentaire et « globale », et qualifient ainsi le procédé de *SRL cross-show* (dénommé SRL de collections dans ce manuscrit). Dans ce contexte, le procédé de SRL réalisé séparément sur chaque

enregistrement d'une collection est alors qualifié de *SRL single-show*, ou tout simplement, SRL (dénommé SRL d'émissions dans ce manuscrit).

Ce chapitre s'organise de la façon suivante : nous évoquerons dans une première partie les travaux fondateurs menés sur l'appariement en locuteurs par [Bonastre et al., 2003b; Meignier et al., 2002], et nous formaliserons le principe de cette tâche en nous appuyant sur le travail de [Leeuwen, 2010]. Nous présenterons dans une seconde partie la relation entre appariement en locuteurs et SRL de collections à l'aide des travaux menés par [Bourlard et al., 2013; Ferràs et Bourlard, 2012; Ghaemmaghami et al., 2013], et nous détaillerons les architectures proposées à cet effet par [Tran et al., 2011; Yang et al., 2011]. Nous profiterons également de cette partie pour évoquer certaines particularités liées à l'évaluation en SRL de collections.

3.1. Appariement en locuteurs

Les premiers travaux pouvant s'apparenter à de la SRL de collections concernent l'appariement en locuteurs *(speaker linking)*. Ces travaux pionniers semblent avoir été menés dans le cadre de la tâche *2-speaker* des évaluations NIST de 2002 en vérification du locuteur [Bonastre et al., 2003b; NIST, 2002]. La tâche de référence en vérification du locuteur, introduite en 1996 et abrégée *1-speaker*, consiste à déterminer si un locuteur cible, pour lequel nous disposons de données d'apprentissage, parle dans un enregistrement mono-locuteur test donné. Une variante dénommée *2-speakers* est introduite en 1999. Dans cette variante, l'enregistrement test correspond à une conversation entre deux locuteurs. Le but est alors de déterminer si le locuteur cible, pour lequel nous disposons de données d'apprentissage (un enregistrement mono-locuteur), correspond à l'un des deux locuteurs de la conversation test.

En 2002, la tâche *2-speakers* est modifiée. L'enregistrement de test correspond toujours à une conversation entre deux locuteurs, en revanche, les données d'apprentissage pour le locuteur cible ne correspondent plus à un unique enregistrement mono-locuteur, mais à trois conversations entre deux locuteurs. Aucune information n'est disponible *a priori* sur ces trois conversations d'apprentissage, si ce n'est qu'elles impliquent exactement deux locuteurs, et que l'un de ces deux locuteurs correspond au locuteur cible (le locuteur cible parle dans les trois conversations). Ainsi, avec la tâche *2-speakers* de 2002, il est nécessaire de déterminer les segments de parole correspondant au locuteur cible, dans les trois conversations, pour l'apprentissage d'un modèle de locuteur cible fiable. Le procédé à mettre en place pour déterminer les segments du locuteur cible correspond à celui de la SRL de collections. Dans un

premier temps, une étape de segmentation et de regroupement en locuteurs doit être effectuée sur chacune des trois conversations, afin de séparer la parole des locuteurs impliqués. Ensuite, il convient de déterminer lequel des locuteurs est commun aux trois conversations.

3.1.1 Les prémices (2002)

L'approche proposée par le Laboratoire d'Informatique d'Avignon (LIA) a permis d'atteindre de très bons résultats sur la tâche *2-speakers* des évaluations NIST de 2002 en vérification du locuteur [Bonastre et al., 2003b]. Cette approche consiste en trois étapes :

1. Une étape de segmentation et regroupement en locuteurs, où les trois conversations d'apprentissage sont traitées séparément de manière à déterminer les segments de parole correspondant aux deux locuteurs impliqués dans chaque conversation.
2. Une étape d'appariement en locuteurs, où les modèles de locuteurs entraînés à partir des segmentations fournies par l'étape 1 sont confrontés afin de déterminer le locuteur cible (qui parle assurément dans chacune des trois conversations d'apprentissage).
3. Une étape de vérification du locuteur, qui vise à déterminer si oui ou non le locuteur cible parle dans la conversation test.

Dans cette partie nous ne détaillerons que l'étape n°2, qui repose essentiellement sur les travaux menés par [Meignier, 2002; Meignier et al., 2002] sur l'indexation en locuteurs. La première étape a été réalisée, dans le cadre des évaluations NIST, avec un système de SRL d'émissions, dont l'architecture et les principales méthodes ont été présentées en chapitre 2. Quant à l'étape n°3, elle ne fera pas l'objet d'une description détaillée du fait de son éloignement avec le sujet.

Dans les travaux menés par [Meignier, 2002; Meignier et al., 2002], l'étape d'appariement en locuteurs (alors dénommé *« speaker tying »*) est réalisée en considérant un *a priori* important : les segmentations fournies par l'étape n°1 sont sans erreur et n'ont pas à être remises en question : elles sont supposées parfaites. Afin de s'affranchir des erreurs de segmentation inhérentes au procédé de SRL d'émissions, les auteurs expérimentent leur approche sur des segmentations de référence fournies par NIST. L'approche proposée repose sur le regroupement agglomératif hiérarchique, et les classes sont représentées par des modèles GMM appris sur les segments de chaque locuteur des segmentations de référence. Ces modèles de locuteur GMM

sont constitués de 128 composantes gaussiennes et ont été entraînés via adaptation MAP d'un modèle du monde. Afin d'estimer la vraisemblance entre deux classes de locuteurs, à chaque itération de l'algorithme de regroupement, les auteurs proposent deux mesures qui utilisent explicitement toutes les informations des enregistrements dont proviennent les deux classes. Ces deux mesures de vraisemblances tirent profit de l'*a priori* établi sur les segmentations utilisées, qui se veulent parfaites. Considérons deux documents A et B. Si un locuteur i du document A et un locuteur j du document B correspondent en réalité à un même locuteur, alors le locuteur i ne peut pas correspondre à d'autres locuteurs du document B, et le locuteur j ne peut pas correspondre à d'autres locuteurs du document A.

Les deux mesures proposées sont présentées dans les équations 3.1 et 3.2, où A_i représente une classe de locuteur i d'un document A, et B_j, une classe de locuteur j d'un document B. $\overline{A_i}$ représente les classes de locuteurs du document A autre que i, et $\overline{B_j}$ représente les classes de locuteur du document B autre que j. La fonction f représente, au choix, la vraisemblance ou le rapport de vraisemblance. $\lambda(A_i)$ (respectivement $\lambda(\overline{A_i})$, $\lambda(B_j)$ et $\lambda(\overline{B_j})$) représente le modèle de locuteur de la classe A_i (respectivement, des classes $\overline{A_i}$, B_j et $\overline{B_j}$).

$$d_1(A_i, B_j) = \frac{f(\overline{B_j}|\lambda(A_i)) + f(\overline{A_i}|\lambda(B_j))}{f(B_j|\lambda(A_i)) \times f(A_i|\lambda(B_j))} \tag{3.1}$$

$$d_2(A_i, B_j) = \frac{f(B_j|\lambda(\overline{A_i})) + f(A_i|\lambda(\overline{B_j}))}{f(B_j|\lambda(A_i)) \times f(A_i|\lambda(B_j))} \tag{3.2}$$

La première des deux mesures (équation 3.1) utilise les données des autres classes que i et j des documents A et B pour établir la similarité entre les classes A_i et B_j. La seconde mesure est similaire, mais utilise les modèles de locuteurs autres que ceux des classes A_i et B_j (c'est-à-dire, $\overline{A_i}$ et $\overline{B_j}$). Afin de s'assurer qu'aucun regroupement ne sera effectué entre les classes de locuteurs provenant d'un même enregistrement, leurs scores de vraisemblance sont artificiellement fixés à $+\infty$ en amont du procédé de regroupement.

Les résultats de l'application des mesures d_1 et d_2 pour la tâche de vérification du locuteur *1-speaker* ont été comparés à ceux obtenus avec la mesure « habituelle » qu'est le rapport de vraisemblance croisé (CLR). Il ressort de ces expériences que la mesure d_1 est plus performante que la mesure CLR lorsque la fonction f correspond au rapport de vraisemblance. En revanche, concernant la tâche d'appariement en locuteur, il apparaît que la mesure de vraisemblance la plus adaptée pour sélectionner les deux classes à regrouper lors d'une itération de l'algorithme de regroupement est la mesure CLR.

Dans le cadre des évaluations NIST en vérification du locuteur, on ne considère que trois conversations (enregistrements téléphoniques de courte durée) à deux locuteurs. Le procédé mis en place est cependant généralisable à n enregistrements contenant chacun plus de deux locuteurs (cas des émissions journalistiques d'information). À noter également que pour la tâche *2-speakers* de l'évaluation NIST de 2002, la mesure de vraisemblance employée pour identifier le locuteur commun aux trois conversations d'apprentissage correspondait au logarithme du rapport de vraisemblance croisé, et l'algorithme de regroupement était fortement contraint, car seulement deux itérations sont nécessaires pour déterminer le locuteur commun [Bonastre et al., 2003b].

3.1.2 Formalisation de la tâche (2010)

[Leeuwen, 2010] propose une étude très complète du problème d'appariement en locuteurs, qu'il définit comme une tâche étroitement liée à la SRL dans laquelle l'objectif est d'identifier les segments de parole, issus d'enregistrements différents, provenant des mêmes locuteurs. D'après cette étude établissant les fondamentaux de la tâche d'appariement en locuteurs, la différence avec la tâche de SRL porte à la fois sur la pluralité des enregistrements (concept de collection) et sur le caractère invariable des regroupements effectués durant la SRL (les segmentations en locuteurs ne sont pas remises en cause, elles sont considérées parfaites, et donc, immuables). Il s'agit d'un problème combinatoire où le nombre de locuteurs est inconnu et peut varier entre 1 et le nombre total de segments en jeu. Le nombre de partitions possibles pour N segments est donné par le N^e nombre de Bell B_N, dont la croissance en fonction de N est très rapide [1] [Brümmer et De Villiers, 2010; Leeuwen, 2010].

Dans cette étude, les principaux aspects du problème sont abordés de manière pertinente à travers des réflexions portant sur :

1. L'aspect dynamique des collections et son implication directe sur le regroupement. Deux types de regroupement hiérarchique sont discernés, selon qu'ils sont réalisés globalement sur l'ensemble des données de la collection (regroupement *off-line*), ou séquentiellement (regroupement *on-line*). Si la quantité de données présente dans la collection venait à augmenter, le regroupement *off-line* devrait alors être à nouveau réalisé sur l'ensemble des données. Son coût global est estimé à $\mathcal{O}(N^4)$ avec l'approche de regroupement agglomératif hiérarchique. Le regroupement *on-line* ne considérerait quant à lui que les nouvelles données de la collection, et son coût global correspond à $\mathcal{O}(C_N)$ avec

1. $B_1 = 1, B_2 = 2, B_3 = 5, B_4 = 15, B_5 = 52, \ldots, B_{10} = 115975, \ldots, B_{20} \approx 5,2 \times 10^{10}, \ldots$

l'approche de regroupement agglomératif hiérarchique, où C_N correspond au nombre de classes présentes pour le traitement du segment N.

2. La ré-estimation des modèles de locuteurs, qui est l'approche généralement adoptée en SRL lorsque deux classes sont regroupées. Habituellement, un nouveau modèle GMM est entraîné sur les données résultant de la fusion de deux classes afin de mieux modéliser la nouvelle classe obtenue. Il s'agit de l'étape la plus coûteuse du procédé de regroupement. Or, cette étape peut être optimisée en ne considérant seulement que la fusion des accumulateurs statistiques des modèles GMM des classes regroupées, permettant ainsi de réduire le coût engendré par la ré-estimation des modèles en évitant un retour aux données.
3. Les mesures d'évaluation. Pour Leeuwen [2010], la tâche d'appariement en locuteurs est indépendante de la SRL. La métrique habituellement utilisée en SRL, le DER, n'est donc pas la plus pertinente. L'auteur propose de se concentrer sur la classification produite à l'issue de l'étape de regroupement, en mesurant l'entropie et l'impureté, à la fois sur les classes et sur les segments de locuteurs.

Si le parallèle est fait avec la tâche de SRL, les expériences menées par Leeuwen [2010] sur l'appariement en locuteurs ne sont cependant pas, à proprement parler, contextualisées par la SRL. D'une part, les expériences menées avec les approches de regroupement en locuteurs reposent sur une matrice symétrique S_{ij} où les scores de vraisemblance entre les différents segments et tous leurs modèles sont établis *a priori* par l'intermédiaire d'un système de vérification du locuteur. D'autre part, les données expérimentales correspondent à des enregistrements téléphoniques provenant de l'évaluation en vérification du locuteur NIST-2006-SRE. En ce sens, il n'est pas surprenant que la métrique DER n'ait pas été retenue pour évaluer les performances de la tâche d'appariement en locuteurs, d'autant plus l'outil d'évaluation officiel fourni par le NIST, qui repose sur des heuristiques, n'est pas capable de gérer des segmentations volumineuses.

La particularité des travaux publiés dans [Leeuwen, 2010] réside essentiellement dans la définition des spécifications pour les regroupements *off-line* et *on-line*. Les algorithmes proposés pour ces deux types de regroupement sont sommaires, mais présentent, toutefois, la particularité de s'affranchir de l'étape coûteuse de réestimation des modèles GMM après chaque regroupement. Les regroupements sont effectués en fonction des scores individuels entre les segments des différents enregistrements. Ces scores, qui sont donc déterminés en amont du procédé de regroupement par un système de vérification du locuteur (matrice symétrique S_{ij}), correspondent à la vraisemblance entre chaque segment et chaque modèle appris à partir de ces mêmes

segments. Le modèle d'une classe C composée de plusieurs segments est estimé simplement à partir des statistiques *Baum-Welch* d'ordre 1 et 0 de ses segments.

▷ Regroupement off-line

Le regroupement *off-line* correspond à l'approche mise en œuvre par [Bonastre et al., 2003b; Meignier et al., 2002], dont les travaux ont été évoqués dans la partie précédente. Il s'agit de l'algorithme de regroupement agglomératif hiérarchique tel que présenté dans la partie 2.3.3. Le regroupement démarre donc avec autant de classes qu'il y a de segments à traiter. Ces classes sont ensuite itérativement regroupées en fonction de la valeur de leurs scores de vraisemblance, jusqu'à ce que le plus élevé des scores soit inférieur à une valeur seuil θ déterminée expérimentalement.

Algorithme 3 : Regroupement *off-line*

1: Chaque segment i est assigné à une classe C_i,
et le nombre de classes $C = N$ (avec N le nombre total de segments).

if *score* s_{max} *le plus élevé* $> \theta$ **then**
  repeat
    2: Trouver le score s_{max} le plus élevé;
    Déterminer les ligne i et colonne j correspondantes dans la matrice S_{ij}
    3: Regrouper les segments des classes C_i et C_j;
    Retirer la colonne j de la matrice S_{ij}
  until *score* $s_{max} < \theta$;
end

▷ Regroupement on-line

Le regroupement *on-line* est envisagé par Leeuwen [2010] comme une approche séquentielle où les segments des différents enregistrements sont distribués dans des classes en fonction des scores de vraisemblance entre les segments et les modèles. L'ordre dans lequel les segments sont traités est fixe. L'algorithme de regroupement *on-line*, présenté ci-dessous, démarre avec une unique classe composée d'un seul segment (le premier segment selon l'ordre établi). S'ensuit alors un procédé itératif dans lequel les segments suivants sont sélectionnés un à un et répartis dans des classes existantes si le score de similarité le permet, ou dans une nouvelle classe le cas échéant. À terme, chaque classe C est considérée comme une collection de segments provenant d'un locuteur en particulier.

Algorithme 4 : Regroupement *on-line*

1: Le premier segment $j = 0$ est assigné à une nouvelle classe C_0,
et le nombre de classes $C = 1$

while *tous les segments ne sont pas assignés à une classe de locuteurs* **do**

2: Avec le segment j suivant, on détermine :
- le score maximum $m_j = \max_{i=0}^{j-1} S_{ij}$
- le modèle i permettant de maximiser m_j

3: Si $m_j > \theta$, le segment j est assigné à la classe contenant
le segment correspondant au modèle i

Sinon, j est assigné à une nouvelle classe C_c,
et le nombre de classes C est incrémenté.

end

Dans l'étude présentée, le seuil de regroupement θ optimal est déterminé empiriquement, et l'ordre dans lequel les segments ont été traités a été déterminé aléatoirement à l'avance.

Les approches de regroupement *on-line* ont déjà fait l'objet d'études dans le domaine spécifique de la SRL d'émissions, en particulier dans [Geiger et al., 2010; Markov et Nakamura, 2007]. En revanche, l'objectif recherché par ces auteurs avec les approches de regroupement *on-line* était la réalisation du procédé de SRL (d'émissions) en temps réel, et non pas l'appariement en locuteurs des segments de différents enregistrements.

3.1.3 Discussion

L'approche de regroupement *on-line* proposée dans [Leeuwen, 2010] consiste à traiter les différents enregistrements de manière séquentielle, et ne se focalise pas sur un traitement en temps réel comme le laisserait supposer la dénomination *on-line*. Nous préférons employer des dénominations différentes pour distinguer ces deux types de regroupements en SRL de collections. Ainsi, dans les parties et chapitres suivants, nous parlerons de « regroupement global » pour désigner le regroupement *off-line*, et de « regroupement incrémental (ou séquentiel) » pour désigner le regroupement *on-line*.

[Leeuwen, 2010] discute sur l'impact de la taille de la collection par rapport aux résultats. Il présente une série de résultats obtenus en faisant varier la quantité de segments impliqués dans le regroupement, et constate que plus le nombre de locuteurs et de segments augmente, plus le problème d'appariement se révèle difficile. Ces

expériences sur l'appariement en locuteurs ont été menées sur des enregistrements téléphoniques entre deux locuteurs (des conversations) issus des données fournies dans le cadre de la campagne d'évaluation NIST en vérification du locuteur de 2006. Il est raisonnable de penser que le constat effectué quant à la difficulté de la tâche d'appariement en locuteur sera d'autant plus exact sur des enregistrements d'émissions journalistiques d'information, qui se veulent plus riches en locuteurs et plus longs en durées.

Une autre discussion intéressante concerne les résultats obtenus par les regroupements *off-line* et *on-line*, qui sont très proches. Il aurait pourtant semblé raisonnable de penser que l'approche *off-line*, disposant d'une vision globale du problème d'appariement, serait plus efficace.

3.2. SRL de collections

Le concept d'appariement en locuteurs a été exploré, plus récemment, par [Bourlard et al., 2013; Ferràs et Bourlard, 2012; Ghaemmaghami et al., 2013; Tran et al., 2011; Yang et al., 2011]. Leurs travaux s'écartent cependant du concept d'appariement en locuteurs tel que formulé par [Leeuwen, 2010] : ils s'inscrivent pleinement dans la tâche de SRL dédiée au traitement des collections de documents audiovisuels, d'où la dénomination « SRL de collections ». Nous présentons dans un premier temps la métrique DER adaptée à l'évaluation de la SRL de collections. Nous détaillerons ensuite les principales architectures de système pour la SRL de collection en nous appuyant des travaux menés par [Tran et al., 2011; Yang et al., 2011]. Nous évoquerons, au passage, les approches mises en œuvre par [Bourlard et al., 2013; Ferràs et Bourlard, 2012; Ghaemmaghami et al., 2013] dans le cadre de leurs expérimentations avec l'approche de regroupement global dite *hybride*.

3.2.1 $\text{DER}_{\text{d'émissions}}$ et $\text{DER}_{\text{de collections}}$

Les travaux menés en SRL de collections sont principalement évalués avec la métrique DER. La définition originale de la métrique DER ne permet pas d'apprécier la qualité des segmentations produites en SRL de collections : la correspondance entre les segmentations générées par le système et les segmentations annotées manuellement est réalisée séparément pour chaque enregistrement. La valeur du DER sur l'ensemble des enregistrements du corpus ne correspond finalement qu'à une moyenne des DER déterminés sur chaque enregistrement. En suivant cette approche

d'évaluation, un locuteur récurrent qui porterait des étiquettes différentes dans les enregistrements évalués ne pénaliserait pas le DER.

Les auteurs des travaux dédiés à la SRL de collections, qui seront présentés dans la suite de chapitre, ont proposé de distinguer deux métriques d'évaluation DER :

1. La première version, dénommée *« single-show DER »* par [Tran et al., 2011; Yang et al., 2011] et *« within-recording DER »* par [Bourlard et al., 2013; Ferràs et Bourlard, 2012; Ghaemmaghami et al., 2013], est identique à la formulation originale et permet d'évaluer les segmentations au niveau *émission* (*i.e.*, les émissions de la collection sont évaluées **séparément**). Dans ce contexte, l'attribution d'étiquettes différentes à un locuteur qui se voudrait récurrent dans plusieurs émissions d'une collection n'aura aucun impact sur le taux d'erreur DER.
2. La deuxième version, dénommée *« cross-show DER »* par [Tran et al., 2011; Yang et al., 2011] et *« across-recording DER »* par [Bourlard et al., 2013; Ferràs et Bourlard, 2012; Ghaemmaghami et al., 2013], permet d'évaluer les segmentations au niveau *collection* (*i.e.*, les émissions de la collection sont évaluées **simultanément**). Ainsi, pour minimiser la valeur du DER sur l'ensemble de la collection, il est nécessaire que les locuteurs récurrents soient identifiés par une étiquette identique et unique dans tous les enregistrements de collection.

Dans ce manuscrit, nous proposons les dénominations $\textbf{DER}_{\textbf{d'émissions}}$ pour désigner la formulation originale de la métrique DER, et $\textbf{DER}_{\textbf{de collections}}$ pour désigner sa généralisation à l'ensemble des enregistrements d'une collection.

3.2.2 Architectures de regroupement global

[Tran et al., 2011; Yang et al., 2011], dans le cadre d'un programme de recherche industrielle sur les technologies d'analyse, de classification et d'utilisation de documents, ont présenté trois différentes architectures de SRL à deux niveaux de regroupement, adaptées au traitement des collections d'enregistrements de type journal d'information. Les approches présentées ont été configurées et évaluées sur des enregistrements de débats télévisés, enregistrés et annotés pour la SRL dans le cadre du programme de recherche concerné. Parmi ces trois architectures, deux proposent un regroupement global : les enregistrements de la collection sont considérés simultanément dans l'étape de regroupement (approches « par concaténation » et« hybride »). La troisième propose un regroupement incrémental (de type *on-line* selon [Leeuwen, 2010]) : les enregistrements sont traités séquentiellement (approche « incrémentale »).

Les données sur lesquelles [Tran et al., 2011; Yang et al., 2011] ont réalisé leurs expériences sont issues d'une émission radio dont certains enregistrements ont été annotés pour la SRL dans le cadre du projet de recherche associé à ce travail. Le corpus de test, dont la durée totale est de 4 heures, est constitué de 23 enregistrements faisant intervenir 49 locuteurs différents. 10 parmi ces 49 locuteurs interviennent dans plusieurs enregistrements.

▷ Approche par concaténation

Dans cette approche de regroupement global, illustrée en figure 3.1, les segmentations des différents enregistrements composant la collection sont regroupées au sein d'un unique et large fichier de segmentation. Cette concaténation [2] est ensuite utilisée comme segmentation d'entrée dans un système de SRL d'émissions, proche de celui décrit dans le chapitre 2.

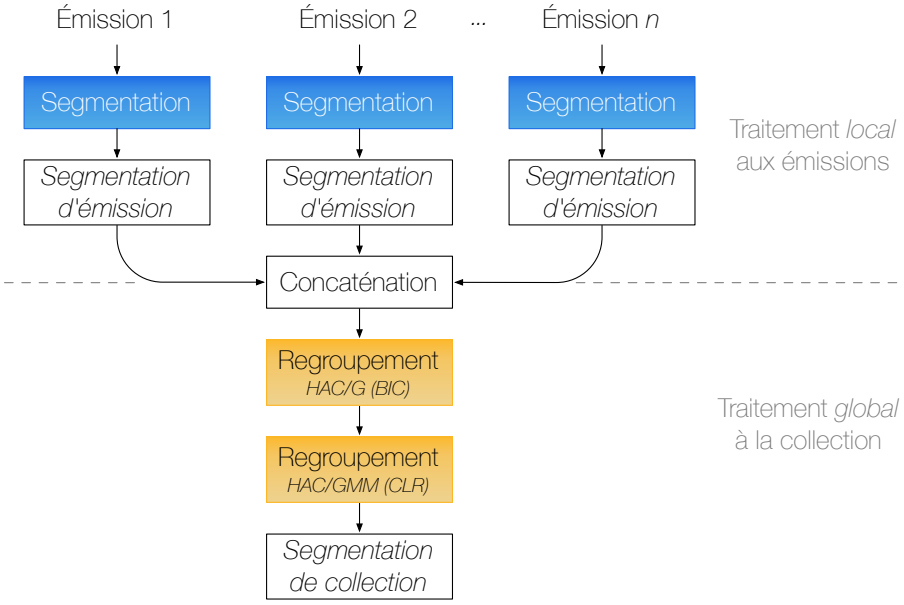

Figure *3.1 – Représentation de l'architecture par concaténation pour la SRL de collections, avec deux étapes de regroupement agglomératif hiérarchique (HAC). Le premier est opéré sur des modèles gaussiens et emploie la mesure BIC, le second, sur des modèles GMM avec la mesure CLR.*

Le premier regroupement agglomératif hiérarchique utilise le critère BIC pour estimer le score de vraisemblance entre les classes, représentées par des gaussiennes à matrices de covariance pleines. Le second regroupement agglomératif hiérarchique profite du résultat du premier regroupement et emploie la mesure CLR pour jauger la proximité des classes, modélisées par des modèles GMM. Il s'agit de la solution la plus

2. Il s'agit du terme employé par les auteurs, bien que l'ordre dans lequel les segmentations sont « réunies » n'ait pas d'importance.

évidente pour mettre en place un système de SRL de collections, ainsi que la moins contraignante. En effet, le système de SRL en soi ne nécessite aucune adaptation particulière, si ce n'est la concaténation des segmentations de chaque émission. Les taux d'erreur produits par le système implémentant cette architecture sont très satisfaisants (de l'ordre de 4% en $\text{DER}_{\text{d'émissions}}$, et 6% en $\text{DER}_{\text{de collections}}$ sur le corpus de *test*). Cependant, cette architecture montre rapidement des limites. La complexité algorithmique du regroupement agglomératif hiérarchique est quadratique, elle augmente en fonction du nombre de classes initiales. Le temps nécessaire au traitement d'une collection est donc fonction de la durée totale des enregistrements qui la compose, or, la capacité mémoire des ordinateurs et autres périphériques n'est actuellement pas suffisante pour exploiter efficacement cette architecture, quelle que soit la taille d'une collection.

▷ Approche hybride

L'approche *hybride*, schématisée en figure 3.2, a été proposée dans le but de résoudre, en partie, le problème de limitation des ressources soulevé dans la définition de l'architecture par concaténation. Étant donné la complexité algorithmique du regroupement agglomératif hiérarchique, il convient de réduire le nombre de classes initiales proposées en entrée.

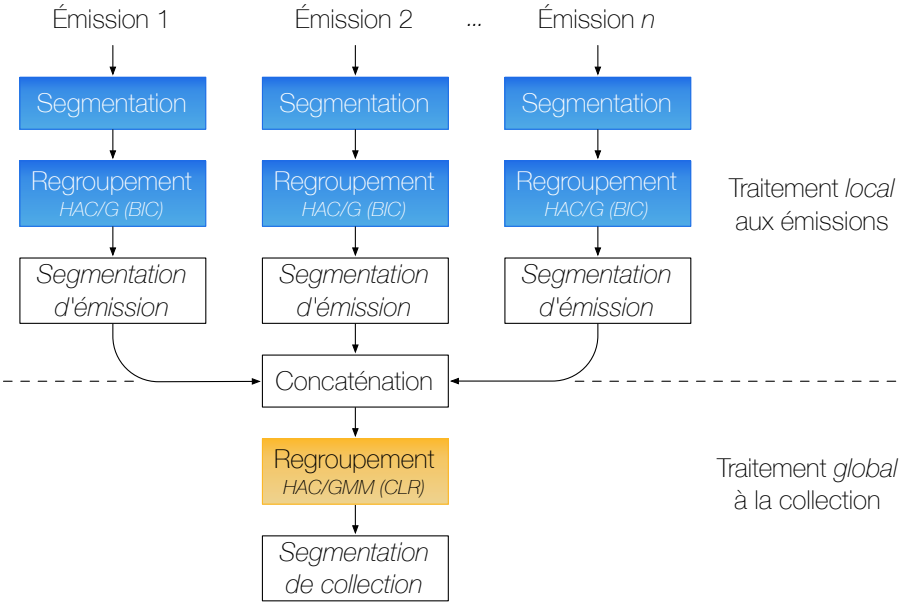

Figure *3.2 – Représentation de l'architecture hybride pour la SRL de collections.*

Pour ce faire, le regroupement faisant appel au critère BIC est réalisé en amont de la concaténation, pour chaque enregistrement de la collection. Le regroupement

utilisant la mesure CLR est donc opéré à partir d'un nombre réduit de classes. Cette approche où le premier regroupement est local aux émissions, et le second, global à la collection (d'où la dénomination *hybride*), ne résout cependant pas le fond du problème. Elle ne fait finalement que le repousser. À collection équivalente, la durée du traitement induite par cette architecture est certes plus courte qu'un traitement par concaténation, cependant, les capacités mémoires ne sont toujours pas suffisantes pour exploiter des collections de tailles conséquentes. En termes de résultats, les taux d'erreur obtenus sont du même ordre que ceux observés avec l'approche par concaténation, suggérant ainsi de privilégier cette approche à la précédente. Les deux architectures présentées jusqu'ici reposent sur le postulat que la collection à traiter est complète (le volume de données n'augmentera pas), et permettent d'estimer le problème de regroupement dans sa globalité. Si toutefois la collection venait à être enrichie par de nouveaux enregistrements, il serait alors nécessaire de :

1. Réaliser le traitement local des nouveaux enregistrements (étape de *segmentation*, ou *segmentation + regroupement BIC*, selon l'architecture),
2. Effectuer à nouveau le traitement global de la collection, en tenant compte des nouvelles segmentations obtenues au niveau local.

[Bourlard et al., 2013; Ferràs et Bourlard, 2012] ont mené des travaux sur des enregistrements de réunions avec une architecture similaire. L'étape d'appariement en locuteurs, qui est effectuée par l'intermédiaire d'un regroupement agglomératif hiérarchique, présente cependant quelques différences. Dans leur approche, le critère de liaison utilisé pour mesurer la distance entre deux classes correspond à la méthode de Ward [Ward Jr, 1963], implémentée de manière récursive avec l'algorithme de Lanc-Williams [Lanc et Williams, 1967]. Les regroupements sont donc effectués en fonction des classes minimisant l'augmentation des variances intra-classes, qui sont déterminées *a priori* sur les hypothétiques regroupements, afin de produire des classes compactes. Les classes de locuteurs sont représentées par des modèles JFA, et plusieurs mesures ont été employées pour estimer la vraisemblance entre les classes, parmi lesquelles nous citerons la similarité cosinus et la version symétrique de la divergence KL.

Les données sur lesquelles [Bourlard et al., 2013; Ferràs et Bourlard, 2012] ont évalué leur approche correspondent aux corpus AMI56 (56 locuteurs, 146 enregistrements, 1044 classes de locuteurs) et AMI56CH (56 locuteurs, 181 enregistrements et 1262 classes de locuteurs). Les meilleurs résultats en termes de DER obtenus par les auteurs sur le corpus AMI56 sont de 21,7% en termes de $\text{DER}_{\text{d'émissions}}$ et 23,6% en termes de $\text{DER}_{\text{de collections}}$. Sur le corpus AMI56CH, ces taux d'erreur augmentent légèrement : 26,8% en $\text{DER}_{\text{d'émissions}}$ et 28,0% en $\text{DER}_{\text{de collections}}$. Ces taux

d'erreur élevés peuvent s'expliquer en partie par la variabilité du locuteur et de l'environnement, qui est très élevé du fait des conditions acoustiques d'enregistrement différentes (plusieurs canaux, plusieurs salles).

[Ghaemmaghami et al., 2013] présente une approche quasiment identique, mais l'applique à des enregistrements d'émissions journalistiques télévisuelles. La différence majeure avec les travaux de [Bourlard et al., 2013; Ferràs et Bourlard, 2012] repose sur le critère de liaison, cette fois maximum, et l'emploi du score de vraisemblance CLR (la vraisemblance des données par rapport au GMM-UBM correspond aux dénominateurs). Les données utilisées par [Ghaemmaghami et al., 2013] pour expérimenter son approche correspondent à 55 enregistrements dont les durées varient d'environ 1 à 6 minutes, ce qui est faible pour des journaux télévisuels. Chaque enregistrement est caractérisé par la présence d'un à neuf locuteurs différents, pour un total de 92 locuteurs sur l'ensemble des enregistrements. L'approche mise en place a permis d'atteindre 13,3% en termes de $\text{DER}_{\text{d'émissions}}$ et 17,0% en $\text{DER}_{\text{de collections}}$.

Les résultats obtenus par [Ghaemmaghami et al., 2013] peuvent difficilement être comparés à ceux obtenus par [Bourlard et al., 2013; Ferràs et Bourlard, 2012], car ni la nature des enregistrements ni le nombre de locuteurs ne sont les mêmes. En revanche, un parallèle peut être fait avec la remarque formulée par [Leeuwen, 2010] quant à la difficulté de la tâche d'appariement qui augmente en fonction du nombre de locuteurs impliqués.

3.2.3 Architecture de regroupement incrémental

[Tran et al., 2011; Yang et al., 2011] proposent également une architecture de regroupement incrémental, qui entre dans la catégorie des regroupements *on-line* tel que formulé par [Leeuwen, 2010]. L'objectif est de s'affranchir des contraintes technologiques sur la capacité des ordinateurs actuels et permettre le traitement des collections dont le volume est susceptible d'augmenter dynamiquement au cours du temps. L'approche incrémentale se démarque par son caractère itératif : contrairement aux deux précédentes architectures, qui offraient une vision globale du problème de regroupement, l'approche incrémentale propose de traiter les émissions les unes après les autres (*cf.* figure 3.3).

Le principe repose sur la propagation des regroupements : on cherche à regrouper les classes provenant du regroupement hiérarchique BIC de l'émission i à celles obtenues lors du regroupement hiérarchique CLR de l'émission $i-1$. Pour ce faire, un module d'identification du locuteur (*Open-Set Identification* – OSI) [Geiger et al., 2010; Markov et Nakamura, 2007] a été inséré entre les regroupements BIC et CLR.

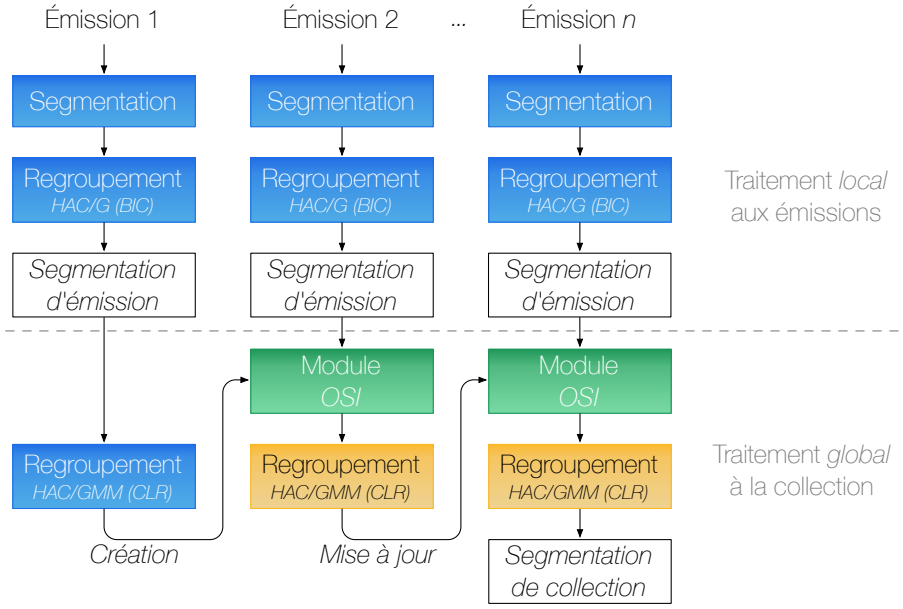

Figure 3.3 – *Représentation de l'architecture incrémentale pour la SRL de collections.*

À l'issue d'une itération i, les modèles GMM représentant les classes fusionnées lors du regroupement hiérarchique CLR sont mémorisés dans ce module OSI. Lors de l'itération $i+1$, les classes proposées en entrée du regroupement hiérarchique CLR correspondent, d'une part, à celles résultant du regroupement hiérarchique BIC de l'émission en cours de traitement (notées c_{BIC}) et, d'autre part, aux classes représentant les modèles GMM mémorisés dans le module OSI (notées c_{OSI}). Le regroupement hiérarchique CLR s'effectue de manière habituelle, cependant, une classe c_{BIC} peut être regroupée avec une autre classe c_{BIC} (cas d'un locuteur n'ayant jamais été détecté par le système), ou avec une classe c_{OSI} (cas d'un locuteur récurrent). Le regroupement s'interrompt lorsque la mesure de vraisemblance entre les classes est inférieure à une certaine valeur déterminée empiriquement. Au terme d'une itération de cette architecture de regroupement, le module OSI est mis à jour pour l'itération suivante :

1. Les modèles GMM des classes c_{OSI} ayant été regroupées avec des classes c_{BIC} sont ré-estimés en fonction des nouvelles données.
2. Les modèles GMM des classes c_{BIC} fusionnées entre elles sont ajoutés au module OSI.

Le premier enregistrement traité ne bénéficie évidemment d'aucune information *a priori*, il permet néanmoins d'initialiser le module OSI. C'est à partir du traitement correspondant au deuxième enregistrement que le module OSI permet *d'identifier* des locuteurs récurrents. Cette approche incrémentale présente deux avantages par

rapport aux alternatives déjà évoquées. Premièrement, elle permet de s'affranchir du problème de capacité mémoire et semble donc tout adaptée au traitement des collections de gros volumes. Toutefois, le nombre de modèles GMM mémorisés par le module OSI augmente en fonction du nombre d'émissions traitées, et donc, la complexité du regroupement hiérarchique CLR également. Cette approche est donc susceptible d'échouer si les collections traitées sont vraiment très conséquentes. Le deuxième avantage réside dans sa capacité à gérer les collections dynamiques, dans lesquelles de nouvelles émissions seraient insérées régulièrement.

En revanche, le $\text{DER}_{\text{de collections}}$ est plus élevé que celui obtenu avec les autres architectures (de l'ordre de 16% en $\text{DER}_{\text{de collections}}$ sur le corpus de *test*, soit +10% en absolu). Cette particularité peut s'expliquer par la propagation des erreurs, évoquée dans la partie 2.3.3, dédiée au regroupement hiérarchique. L'approche incrémentale multiplie les regroupements hiérarchiques CLR là où les approches par concaténation, et hybride, n'en réalisent qu'un seul. Le regroupement hiérarchique fusionne itérativement les classes maximisant la valeur de vraisemblance afin d'approcher un regroupement global optimal. Or, en raison de sa nature, l'approche incrémentale ne dispose pas d'une vision globale du problème. Les auteurs démontrent expérimentalement cet inconvénient en modifiant l'ordre de traitement des émissions de la collection, faisant ainsi varier les $\text{DER}_{\text{d'émissions}}$ et $\text{DER}_{\text{de collections}}$.

3.2.4 Discussion

Les récents travaux menés par [Bourlard et al., 2013; Ferràs et Bourlard, 2012; Ghaemmaghami et al., 2013; Tran et al., 2011; Yang et al., 2011] complètent les travaux initialement réalisés pour les évaluations NIST par [Leeuwen, 2010; Meignier et al., 2002].

Dans ces travaux sur la SRL de collections, la tâche d'appariement en locuteurs est considérée comme extension de la tâche de SRL d'émissions. Les approches mises en œuvre par les différents auteurs pour établir la correspondance entre les locuteurs de différents enregistrements reposent sur le regroupement agglomératif hiérarchique. La différence entre leurs travaux porte principalement sur le type d'enregistrement et leur quantité, ainsi que sur les approches de modélisations employées (qui n'ont cessé d'évoluer depuis le début des années 2000, *cf.* tableau 2.1 sur l'évolution des techniques de modélisation en locuteurs). La métrique d'évaluations employée pour mesurer la qualité des segmentations produites par les systèmes de SRL de collections repose sur une généralisation du DER.

L'approche de regroupement la plus plébiscitée pour effectuer la SRL de col-

lections, mais aussi la plus intuitive, est le regroupement global. Nous présentons dans le tableau 3.1 un récapitulatif des résultats obtenus en SRL de collections avec l'approche de regroupement global sur des collections de différentes natures. Ces résultats ne sont pas comparables entre eux, mais permettent néanmoins de constater que plus le nombre d'enregistrements augmente, et donc plus le nombre de locuteurs à classifier est élevé, plus les $DER_{\text{de collections}}$ se détériorent.

Année	Collection	Genre	n^{bre} enr.	n^{bre} loc.	$DER_{\text{de collections}}$
2011	The Naked Scientists	Radio (débats)	23	[49 ; 10]	6,1%
2012	AMI8	Meetings	18	[135 ; 8]	8,5%
2012	AMI56	Meetings	56	[1044 ; 56]	23,6%
2012	AMI856CH	Meetings	85	[1262 ; 56]	28,0%
2013	SAIVT-BNEWS	TV (journaux)	55	[92 ; -]	17,0%

Table *3.1 – Résultats en termes de* **$DER_{\text{de collections}}$** *obtenus par* **regroupement global** *sur différents* corpus *en SRL de collections. Le nombre de locuteurs des collections est exprimé sous le formalisme suivant : [n^{bre} total ; n^{bre} récurrents].*

Année	Collection	Genre	n^{bre} enr.	n^{bre} loc.	$DER_{\text{de collections}}$
2011	The Naked Scientists	Radio (débats)	23	[49 ; 10]	14,8%

Table *3.2 – Résultats en termes de* **$DER_{\text{de collections}}$** *obtenus par* **regroupement incrémental** *en SRL de collections. Le nombre de locuteurs des collections est exprimé sous le formalisme suivant : [n^{bre} total ; n^{bre} récurrents].*

[Tran et al., 2011; Yang et al., 2011] sont les seuls à avoir expérimenté l'approche de regroupement incrémentale pour la SRL de collection. Les résultats obtenus sont clairement moins bons qu'avec l'approche de regroupement global (14,8% contre 6,1% en $DER_{\text{de collections}}$), cependant, le procédé mis en œuvre permet de traiter des collections dont le volume augmente dynamiquement.

3.3. Bilan général sur la SRL de collections

Plusieurs questions se posent au terme de cet état de l'art sur la SRL de collections.

Tout d'abord, par rapport au constat d'abord effectué par [Leeuwen, 2010] quant à la difficulté de la tâche d'appariement en termes de résultats de classification. L'auteur constate suite à ses expériences sur des *collections* de différentes tailles que plus la collection est fournie, plus le nombre de locuteurs augmente et plus la qualité de la classification en locuteurs se détériore. La question qui se pose alors concerne l'évolution des résultats de classification en SRL de collections si le volume d'enregistrements traité est plus conséquent. Si la détérioration des résultats est fonction

de la taille des collections traitées, comment faire pour gérer des collections dont le volume horaire correspondrait à plusieurs dizaines d'heures et plusieurs centaines ou milliers de locuteurs ?

Concernant les approches mises en œuvre pour effectuer la SRL de collections, deux types de classifications ont été définies et expérimentées : les regroupements de type global (général, *off-line*), et les regroupements de type incrémental (ou séquentiel, itératif, *on-line*). Les expériences menées sur le regroupement global sont nombreuses, et toutes ont été réalisées à l'aide de l'algorithme de regroupement agglomératif hiérarchique. Cette approche de regroupement permet d'obtenir de bons résultats en classification du locuteur, mais son coût algorithmique peut être élevé si de nouveaux modèles de locuteurs sont calculés pour les classes regroupées. Si ce procédé est de durée raisonnable en SRL d'émissions, il peut être fortement contraignant en SRL de collections lorsque le volume d'enregistrements traité est conséquent. [Tran et al., 2011] considère que le regroupement global est efficace, mais peu réaliste d'un point de vue applicatif. Cette réflexion rejoint la précédente sur la qualité de résultats de classification : les approches de regroupement actuelles, et les techniques de modélisation qui vont avec, sont-elles suffisamment efficaces pour gérer le traitement de collections volumineuses ?

La partie suivante, qui présente les contributions apportées durant cette thèse, tente de répondre à ces questions sur la SRL de collections volumineuses. Nous y présentons, dans un premier chapitre, les données manipulées et les collections étudiées. Dans un deuxième et troisième chapitre, respectivement consacré à l'approche de regroupement global et à l'approche de regroupement incrémental, nous présentons les différentes architectures et approches de regroupement étudiées et discutons les résultats obtenus sur les collections étudiées.

Deuxième partie

SRL pour les collections volumineuses

CHAPITRE 4

Présentation des données expérimentales

Ce chapitre, le premier de cette partie dédiée à la présentation des contributions apportées durant cette thèse, porte sur la présentation des données expérimentales utilisées dans le cadre de cette thèse. Ces données correspondent essentiellement aux enregistrements des émissions télévisuelles ayant servi dans la campagne d'évaluation ETAPE et dans le défi REPERE. Ces deux projets, bien que différents, ont été supportés, financés et organisés par les mêmes organismes.

Dans ce chapitre, nous introduisons en premier lieu la campagne d'évaluation ETAPE et le défi REPERE. Nous y présentons les données délivrées dans le cadre de ces deux évènements, puis nous présentons les collections arbitrairement constituées pour expérimenter les approches que nous proposons dans les chapitres suivants.

4.1. Introduction à la campagne d'évaluation ETAPE

La campagne d'évaluation ETAPE [Gravier et al., 2012] (Évaluations en Traitement Automatique de la ParolE) fait suite aux campagnes ESTER [Galliano et al., 2005, 2009] (Évaluation des Systèmes de Transcription d'Émissions Radiophoniques) qui furent organisées en 2003, 2005 et 2009. Ces dernières visaient à évaluer les performances des systèmes de transcription sur des émissions radiophoniques d'information. La difficulté reposait alors sur le type de parole traité (parole lue/préparée et parole spontanée) et la qualité des enregistrements (environnement bruité ou non). ETAPE proposait d'étendre cette tâche à des enregistrements de qualité sonore et de sources variées, en proposant d'utiliser des enregistrements audiovisuels dont le degré de parole spontanée et de parole superposée était variable. Cette campagne d'évaluation, qui a débuté en mars 2011 et s'est clôturée en mai 2012, a été partiellement

financée par l'Agence Nationale de la Recherche (ANR). Les principaux partenaires de ce projet étaient l'Association Francophone de la Communication Parlée (AFCP), la Direction Générale de l'Armement (DGA), l'agence ELDA (Evaluations and Language resources Distribution Agency) et le Laboratoire National de métrologie et d'Essais (LNE). Quatre tâches indépendantes ont été évaluées, dont la tâche de segmentation et regroupement en locuteurs. ETAPE présentait la particularité, dans la tâche de SRL, de considérer deux variantes, selon que la détection des tours de parole était locale aux émissions (SRL), ou globale à l'ensemble des émissions du corpus (SRL-X).

Les données fournies pour participer à la campagne ETAPE correspondent à environ 14 heures d'enregistrements radiophoniques et 29 heures d'enregistrements audiovisuels. La répartition des données, en termes de durée audio et durée évaluée (calculées à l'aide des fichiers UEM [1] fournis par les organisateurs), en fonction du genre des émissions, est présentée dans le tableau 4.1.

Genre	Durée Audio	Durée UEM	Émissions
Journaux télévisés	10h47	7h46	*BFM Story* (BFMTV) *Top Questions* (LCP)
Débats politiques	15h50	13h02	*Pile et Face* (LCP) *Ça vous Regarde* (LCP) *Entre les Lignes* (LCP)
Divertissements télévisés	2h15	1h34	*La Place du Village* (TV8)
Émissions radiophoniques	14h13	11h27	*Un Temps de Pauchon* *Service Public* *Le Masque et la Plume* *Comme on nous Parle* *Le Fou du Roi*
Total	**43h05**	**33h49**	

Table *4.1 – Répartition des données ETAPE, en termes de durée audio et durée évaluée en fonction du genre des émissions.*

4.2. Introduction au défi REPERE

Le défi REPERE (REconnaissance de PERsonnes dans des Émissions audiovisuelles) est un projet d'évaluations dans le domaine de la reconnaissance multimédia de personnes dans des documents télévisuels [Galibert et Kahn, 2013]. Ce projet, qui a débuté en mars 2011 et s'est clôturé en mars 2014, a été cofinancé par l'ANR et

1. Unpartitioned Evaluation Map, index des segments audio à évaluer.

la DGA, et co-organisé par le LNE et l'agence ELDA. La tâche principale consiste à identifier les personnes présentes dans les émissions, qu'elles apparaissent à l'écran ou qu'elles interviennent oralement. La tâche secondaire vise à déterminer les personnes mentionnées dans les émissions, dont l'identité a été affichée à l'écran, ou citée oralement. La tâche de segmentation et regroupement en locuteurs correspond à l'une des sous-tâches évaluées en marge des tâches principales.

Ce projet, qui repose essentiellement sur la combinaison d'informations extraites à partir des modalités audio et vidéo, nécessite des compétences variées. Les participants impliqués dans ce projet – laboratoires universitaires et industriels – se sont regroupés en consortiums afin de gérer cet aspect multimodal, et ainsi couvrir l'ensemble des thèmes proposés par le défi REPERE. Le LIUM et l'institut de recherche suisse IDIAP ont associé leurs compétences en traitement automatique de la parole et traitement vidéo pour constituer le consortium SODA (reconnaiSsance de persOnnes pour les Débats et les journAux télévisés). Le Centre de Recherche en Informatique de Montréal (CRIM) s'est joint au consortium SODA en 2013, diversifiant ainsi les approches proposées en traitement automatique de la parole. Deux autres consortiums ont participé à ce projet d'évaluation : QCOMPERE et PERCOL [Galibert et Kahn, 2013].

Le défi REPERE a été organisé en trois phases, chacune étant accompagnée d'un nouvel ensemble de données, et ponctuée par une évaluation. Les évaluations ont eu lieu en janvier 2012 *(Dry-run)*, 2013 (Phase 1) et 2014 (Phase 2). Les données sélectionnées par ELDA correspondent à des émissions enregistrées sur les chaînes de télévision française d'information BFMTV et LCP, accessibles par la télévision numérique terrestre française.

Le corpus REPERE est donc constitué de l'ensemble des corpora d'apprentissage, de développement et de test des Phases 1 et 2. Le corpus REPERE représente environ 169 heures de données audiovisuelles, réparties sur 360 émissions enregistrées entre mars 2011 et avril 2013. La répartition de ces données, en termes de durée audio, est présentée dans le tableau 4.2. Seulement 60 heures de l'ensemble du corpus ont été annotées à des fins d'évaluation, les durées des segments évalués sont présentées dans la colonne *durée UEM*.

Les enregistrements qui constituent le corpus REPERE correspondent à des émissions de différents types, enregistrées sur les chaînes de télévision française BFMTV et LCP:

- *Top Questions (LCP)* est une émission dans laquelle les députés interrogent les ministres sur l'actualité (type « questions au gouvernement »).

Genre	Durée audio	Durée UEM	Émissions
Débats politiques	51h32	15h59	*Ça vous Regarde* (LCP) *Pile et Face* (LCP) *Entre les Lignes* (LCP)
Journaux télévisés	67h52	32h16	*BFM Story* (BFMTV) *Ruth Elkrief* (BFMTV) *LCP Info* (LCP) *LCP Actu* (LCP)
Questions au gouvernement	11h09	6h32	*Top Questions* (LCP)
Magazine people	38h36	5h00	*Culture et Vous* (BFMTV) *Planète Showbiz* (BFMTV)
Total	**169h09**	**59h47**	

Table *4.2 – Répartition des données du défi REPERE, en termes de durée audio et de durée évaluée.*

- *Ça vous Regarde (LCP)*, *Pile et Face (LCP)* et *Entre les Lignes (LCP)* correspondent principalement à des débats sur l'actualité et la politique.
- *LCP Info (LCP)*, *LCP Actu (LCP)*, *BFM Story (BFMTV)* et *Ruth Elkrief (BFMTV)* correspondent à des journaux d'information modernes relatant et discutant l'actualité quotidienne.
- *Culture et Vous (BFMTV)*, anciennement *Planète Showbiz*, est une émission portant sur l'actualité des personnalités et célébrités.

4.3. Découpage en collections

Nous avons utilisé les données fournies dans la campagne ETAPE et le défi REPERE pour constituer des collections. Plus précisément, nous avons considéré l'ensemble des émissions BFMTV et LCP des corpus de développement et de test de la campagne ETAPE, ainsi que l'intégralité des données fournies lors du défi REPERE. Les données du corpus d'apprentissage de la campagne ETAPE ont été utilisées pour l'apprentissage des modèles, par conséquent elles n'ont pas été sélectionnées. Nous sommes donc en présence d'une collection de 374 enregistrements, couvrant une période de 26 mois (de février 2011 à avril 2013) dont la durée totale est d'environ 178 heures d'audio.

4.3.1 Collections d'émissions

Pour rappel, les collections d'émissions regroupent des enregistrements présentant une ou plusieurs caractéristiques communes, et plusieurs niveaux de collection

d'émissions peuvent être envisagés. Les enregistrements fournis dans le cadre de la campagne ETAPE et du défi REPERE permettent de travailler sur trois différents niveaux de granularité (*cf.* figure 4.1):

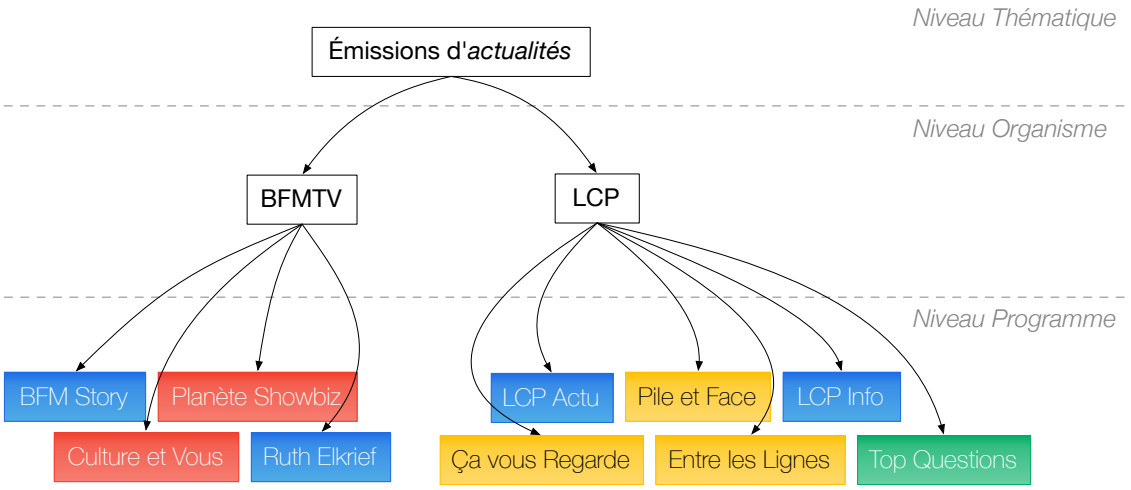

Figure *4.1 – Représentation hiérarchique des collections d'émissions par niveau d'étude. Les couleurs représentent le* type *des émissions.*

Les émissions présentes dans les corpus ETAPE et REPERE permettraient également de s'intéresser aux collections d'un point de vue typologique. Par exemple, les émissions *Ça vous Regarde*, *Pile et Face* et *Entre les Lignes* constituent une collection dont les enregistrements concernent essentiellement des débats politiques (en jaune sur la figure 4.1). Ce contexte d'analyse n'est cependant pas abordé dans le cadre de cette thèse. Les émissions *Top Questions (LCP)* et *Culture et Vous – Planète Showbiz (BFMTV)*[2] constituent à elles seules des types d'émissions isolés (respectivement « questions au gouvernement » et « magazine *people* »), et l'analyse de ces émissions, en terme de SRL, est déjà couverte au niveau *programme*. Le tableau 4.3 présente les 10 collections d'émissions étudiées dans cette thèse (7 collections *programme*, 2 collections *organisme*, 1 collection *thématique*). Pour chaque collection sont présentés : le nombre d'enregistrements qui les constitue, la durée totale de la modalité audio et la durée évaluée (qui est précisée dans les fichiers UEM associés), le nombre total de locuteurs et le nombre de locuteurs récurrents (qui interviennent dans au moins deux émissions de la collection).

Des regroupements d'enregistrements ont été effectués pour constituer trois collections en particulier. Le corpus REPERE n'inclut que quatre enregistrements *Ruth Elkrief*, et seulement trois enregistrements *LCP Actu*. Les enregistrements de l'émission *Ruth Elkrief* ont été regroupés avec ceux de l'émission *BFM Story*, diffusée

2. *Planète Showbiz* et *Culture et Vous* désignent une même émission dont le nom a changé au cours de la campagne REPERE.

Niveau	Collection		nbre enr.	Durée audio	Durée UEM	nbre locuteurs total	nbre locuteurs récur.
Programme	BFMTV	**BFM Story** *(+ Ruth Elkrief)*	48	49h32	23h00	556	83
		Planète Showbiz *(+ Culture et Vous)*	160	38h27	5h00	771	75
	LCP	**Ça vous Regarde**	23	20h55	7h14	173	11
		Entre les Lignes	27	16h14	7h05	18	9
		LCP Info *(+ LCP Actu)*	48	20h34	11h14	317	96
		Pile et Face	33	19h44	6h11	46	15
		Top Questions	35	12h20	7h28	119	36
Organisme	**BFMTV**		208	88h00	28h00	1309	160
	LCP		166	89h48	39h13	544	172
Thématique	**BFMTV + LCP**		374	177h48	67h13	1787	333

Table *4.3 – Nombre d'enregistrements, durée totale de la collection et durée évaluée (UEM), nombre de locuteurs total et récurrents, pour chacune des collections d'émissions étudiées.*

sur la même chaîne de télévision (BFMTV), en raison de leur proximité en termes de contenu. Un regroupement similaire a été établi avec les enregistrements des émissions *LCP Actu* et *LCP Info*.Les émissions *Planète Showbiz* et *Culture et Vous* désignant une même émission dont le nom a changé au cours de la campagne REPERE, leurs enregistrements sont regroupés au sein d'une seule et même collection.

4.3.2 Collections temporelles

Les collections temporelles sont constituées d'enregistrements de nature hétérogènes couvrant une période bien déterminée, correspondant par exemple à la médiatisation d'un évènement particulier. Étudier ce type de collection en SRL devrait permettre de révéler les différents acteurs impliqués durant l'évènement ciblé. Un problème se pose cependant quant au découpage de nos données en collections temporelles. Les données fournies durant les campagnes ETAPE et REPERE ne sont pas annotées par rapport aux sujets abordés, il n'était donc pas aisé de réaliser un découpage orienté évènements médiatiques. Il nous a pourtant semblé nécessaire d'étudier la SRL de collections sous cette perspective, nous avons donc réparti les enregistrements selon leurs dates de diffusions, en considérant un tri par ordre chronologique.

D'une manière générale, les enregistrements des différentes émissions recueillis pour les campagnes d'évaluation ETAPE et REPERE on été enregistrés sur plusieurs jours consécutifs. On constate toutefois la présence de certaines irrégularités dans la fréquence des enregistrements. Ces irrégularités n'ont rien de surprenant compte tenu

du fait que les données proviennent de deux campagnes d'évaluation différente, et que le défi REPERE a été fractionné en trois évaluations. Si d'une manière générale nous disposons d'au moins un enregistrement par jour, en considérant le tri chronologique des enregistrements, il arrive que deux enregistrements consécutifs soient espacés de plusieurs jours. La durée de ces irrégularités, ou *interruptions* d'acquisition, varie de 2 à 54 jours.

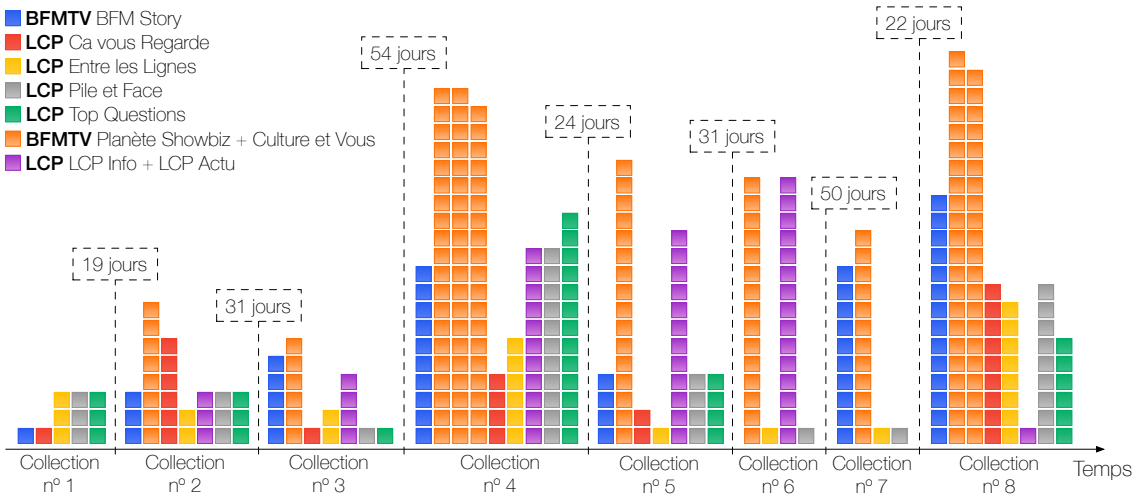

Figure *4.2 – Représentation schématique de la constitution des collections temporelles, avec la durée des interruptions d'enregistrement en nombre de jours. Chaque collection est représentée par la quantité et la répartition des enregistrements qui la compose.*

Nous avons utilisé les interruptions d'enregistrements supérieures à 15 jours pour définir des collections temporelles. Ainsi, tous les enregistrements consécutifs pour lesquels la durée écoulée entre deux enregistrements est inférieure à 15 jours sont regroupés au sein d'une même collection. Cette répartition des données, présentée schématiquement en figure 4.2, permet ainsi de définir 8 collections temporelles plus ou moins volumineuses. Le détail de 8 collections, en termes de durées et de nombre de locuteurs, est présenté ci-dessous dans le tableau 4.4. Les valeurs [x ; y ; z] de la colonne rendant compte du nombre de locuteurs dans une collection correspondent respectivement au nombre total de locuteurs, au nombre de locuteurs *récurrents* et au nombre de locuteurs *récurrents* qui interviennent dans au moins deux enregistrements provenant d'émissions différentes.

Cette répartition arbitraire des données en collections temporelles a permis de constituer des collections de dimensions variables en durées et en nombre de locuteurs. La proportion de locuteurs récurrents par rapport au nombre de locuteurs total est relativement stable, quelle que soit la collection temporelle : entre 11,9% et 20% des locuteurs sont récurrents. Avec les collections d'émissions, définies et présentées dans la partie précédente, cette proportion varie de 6,4% à 32,6% selon la collec-

Collection	Période couverte	n^{bre} enr.	Durée		n^{bre} locuteurs
			audio	UEM	
Temporelle n°1	73 jours	11	6h11	3h52	[58 ; 10 ; **2**]
Temporelle n°2	30 jours	29	15h42	5h46	[217 ; 26 ; **9**]
Temporelle n°3	9 jours	20	10h20	2h45	[140 ; 28 ; **14**]
Temporelle n°4	105 jours	114	39h44	21h01	[669 ; 108 ; **42**]
Temporelle n°5	51 jours	46	18h02	6h26	[247 ; 51 ; **17**]
Temporelle n°6	44 jours	32	9h29	4h19	[205 ; 30 ; **0**]
Temporelle n°7	26 jours	24	12h36	8h38	[262 ; 37 ; **0**]
Temporelle n°8	121 jours	90	61h55	13h21	[461 ; 64 ; **27**]

Table *4.4 – Période couverte en nombre de jours, nombre d'enregistrements, durée totale et durée évaluée (UEM), nombre de locuteurs (formalisme [n^{bre} locuteurs total ; n^{bre} locuteurs récurrents ; n^{bre} locuteurs récurrents sur des enregistrements provenant d'émissions différentes]), pour chacune des collections temporelles étudiées.*

tion. La répartition des enregistrements en fonction de leurs dates de diffusion, en se basant sur les *interruptions* dans la fréquence d'acquisition des enregistrements, semble ainsi permettre la constitution de collections plus homogènes en termes de locuteurs récurrents.

Il aurait été plus intéressant de constituer les collections temporelles en fonction des évènements médiatiques couverts dans les différents enregistrements, néanmoins, la méthode de répartition choisie semble correspondre aux attentes énoncées. En effet, la plupart des collections temporelles ainsi définies incluent des locuteurs récurrents intervenant dans des enregistrements d'émissions différentes (nombre en gras dans la dernière colonne du tableau 4.4). Les collections temporelles n° 6 et 7 n'incluent malheureusement pas ce genre de locuteurs récurrents. Cette particularité peut s'expliquer par la faible diversité d'émissions qui compose ces deux collections (*cf.* illustration 4.2).

4.4. Annotation des données

Les performances des systèmes de SRL sont habituellement évaluées par comparaison entre une segmentation fournie par le système et la segmentation de référence annotée manuellement (métrique DER). Dans le cadre de la SRL de collections, il est primordial que les annotations en locuteurs aient été rigoureusement vérifiées. Les segments d'un même locuteur doivent être identifiés par une étiquette unique et strictement identique au sein de tous les enregistrements d'une collection. Les données distribuées durant les dernières campagnes d'évaluation française ETAPE et REPERE ont été annotées en conséquence, permettant ainsi d'évaluer par collec-

tions.

Ce travail d'annotation est généralement réalisé par plusieurs personnes différentes, ce qui multiplie les risques d'erreur. Un même locuteur dont le prénom serait orthographié de deux manières différentes dans les segmentations de référence, par exemple, *Jamel_Debbouze* et *Djamel_Debbouze*, provoquerait une erreur de confusion lors du calcul de la métrique DER si dans la segmentation fournie par le système les segments correspondants portent la même étiquette (à juste titre).

4.5. Bilan

Nous avons présenté, dans ce chapitre les différentes collections sur lesquelles seront évaluées les approches que nous présentons pour la SRL de collections volumineuses. Ces collections, dont les durées totales sont très hétérogènes (de 6 à 177 heures de données), constituent un matériel convenable pour étudier le comportement d'un système de SRL de collections. En effet, la plus réduite de nos collections, la collection temporelle n°1, est déjà plus fournie en locuteurs et de plus longue durée que les collections étudiées dans le contexte d'émissions journalistiques d'information par [Ghaemmaghami et al., 2013; Tran et al., 2011; Yang et al., 2011].

CHAPITRE 5

SRL de collections par regroupement global

Ce deuxième chapitre sur les contributions apportées durant la réalisation de cette thèse porte sur le regroupement de type *off-line*, ou comme nous préférons l'appeler, le regroupement *global* à la collection. Les travaux présentés dans ce manuscrit ont été contextualisés par la participation du LIUM à la campagne d'évaluation ETAPE et au défi REPERE, qui ont été présentées dans le chapitre précédent. Ces deux évènements ont procuré un cadre de travail favorable à l'étude des collections dans un contexte de SRL, en proposant notamment des évaluations officielles en SRL de collections (les données fournies ont donc été annotées en conséquence). Le point de départ des travaux que nous présentons repose sur la publication des travaux de [Tran et al., 2011; Yang et al., 2011], sur les architectures de SRL adaptées au traitement des collections d'enregistrements, présentés dans le chapitre 3, ainsi que celle sur le regroupement ILP et la modélisation i-vector [Rouvier et Meignier, 2012], présentée dans le chapitre 2.

Ce chapitre, qui présente les travaux menés en SRL de collections sur la base d'une architecture de regroupement global, s'organise de la façon suivante. Nous présentons dans un premier temps l'architecture pour le regroupement global des collections que nous avons mise en place. Nous présentons ensuite une comparaison des différentes méthodes de classification expérimentées, ainsi que les différentes améliorations apportées afin d'optimiser notre approche de regroupement global pour le traitement des collections volumineuses. Nous proposons une architecture et une configuration adaptées, une simplification du problème de regroupement par ILP, ainsi qu'une réduction de sa complexité par la théorie des graphes et, finalement, une étude sur la liberté laissée au système de regrouper globalement les classes intrinsèques aux enregistrements. Nous proposons, dans une troisième partie, une synthèse de l'évaluation de notre système de SRL de collections par regroupement global sur les

différentes collections construites à partir des données ETAPE et REPERE, telles que présentées dans le chapitre précédent.

5.1. Architecture pour le regroupement global

L'architecture pour le regroupement global que nous proposons dans ce manuscrit repose sur les conclusions que nous avons tirées d'une série d'expériences préliminaires reportées en annexe C (p.195). [Tran et al., 2011] ont proposé trois architectures pour le traitement des collections (*cf.* partie 3.2) : deux architectures de regroupement global (par *concaténation* et *hybride*) et une architecture de regroupement incrémental. Cette dernière est étudiée en détail dans le chapitre suivant. Les auteurs concluent que l'approche hybride permet d'obtenir des taux d'erreur comparables à l'approche par concaténation globale, tout en minimisant la durée du traitement. L'architecture que nous proposons pour la SRL de collections par regroupement global repose donc sur l'approche hybride, caractérisée par deux niveaux de traitements indépendants :

1. Le niveau « émission », ou *local*, où chaque enregistrement de la collection est traité séparément avec un système de SRL d'émissions. Le système produit donc des segmentations spécifiques à chaque enregistrement. À ce niveau, les locuteurs communs à plusieurs enregistrements ne sont pas détectés.
2. Le niveau « collection », ou *global*, qui consiste en un regroupement sur l'union des segmentations produites au niveau local. Ce regroupement vise à détecter les locuteurs présents dans plusieurs enregistrements, en leur attribuant une étiquette de locuteur unique à la collection.

Notre version de l'architecture hybride, illustrée en figure 5.1, diffère quelque peu de celle proposée par [Tran et al., 2011]. Nous avons constaté, lors de nos travaux préliminaires, que le meilleur moyen de minimiser le $\text{DER}_{\text{de collections}}$ consiste à produire des segmentations locales de bonne qualité. Dans notre approche, le traitement local aux émissions est réalisé par un système de SRL d'émissions à l'état de l'art, là où [Tran et al., 2011] ne propose qu'un regroupement hiérarchique utilisant le critère BIC. Ce système à l'état de l'art est comparable à celui présenté dans [Rouvier et Meignier, 2012], dans lequel la dernière étape de regroupement est réalisée au moyen de la méthode de classification ILP avec des classes modélisées par l'approche i-vector. Les segmentations produites par ce système sont donc supposées être de très bonne qualité au niveau *émission*.

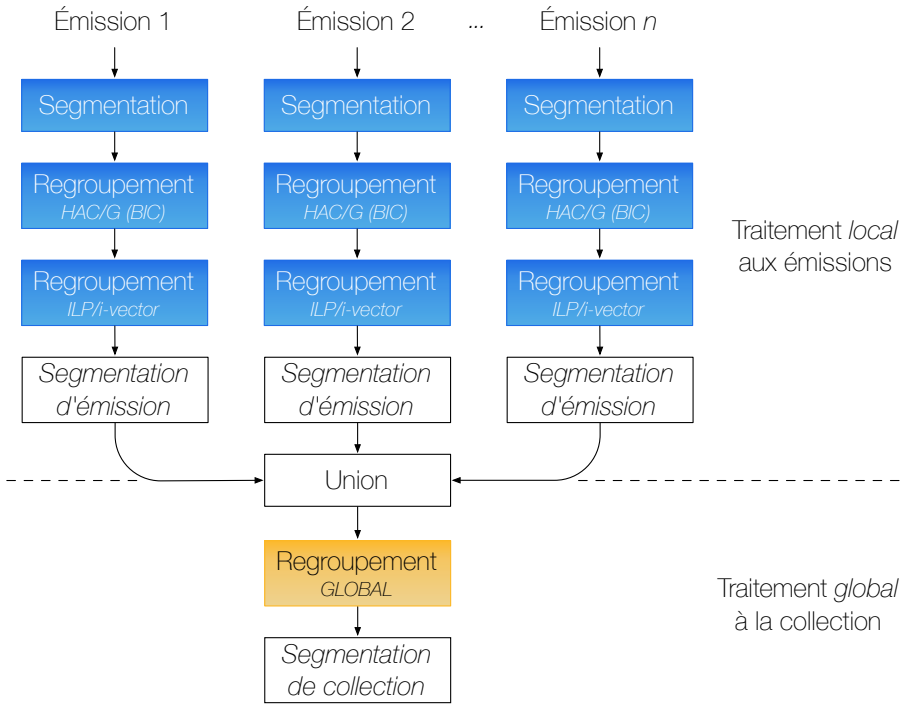

Figure *5.1 – Architecture de regroupement global pour la SRL de collections incluant un système de SRL d'émissions à l'état de l'art pour le traitement local aux émissions.*

Ces segmentations locales sont alors réunies au sein d'un unique fichier de segmentation. Dans cette volumineuse segmentation, les étiquettes attribuées aux différents segments de parole doivent permettre d'identifier les enregistrements dont ils proviennent. Ces étiquettes sont donc rendues uniques pour chaque couple (classe de locuteur, enregistrement de provenance). Le traitement global à la collection consiste en un regroupement portant sur la réunion des segmentations locales aux émissions. Deux stratégies différentes peuvent être envisagées pour ce regroupement global, selon les hypothèses formulées quant aux segmentations du niveau *émission* de l'architecture :

1. Si l'on considère que les segmentations provenant du niveau *émission* sont optimales, alors les regroupements ne peuvent se faire qu'entre des classes provenant d'enregistrements différents, afin de ne pas altérer les segmentations du niveau *émission*. Par conséquent, les $\text{DER}_{\text{d'émissions}}$ obtenus sur les segmentations du niveau *collection* sont identiques à ceux obtenus sur les segmentations du niveau *émission*.

2. Si l'on considère que les segmentations du niveau *émission* sont imparfaites, alors il est possible de laisser la liberté au système de regrouper des classes provenant d'un même enregistrement. De ce fait, les $\text{DER}_{\text{d'émissions}}$ sont susceptibles d'évoluer entre les segmentations des niveaux *émission* et *collection*.

Dans les sections suivantes, nous présentons les méthodes de classification expérimentées pour le regroupement global, ainsi que les améliorations apportées. La stratégie de regroupement suivie pour présenter ces approches est celle laissant la liberté au système de regrouper des classes provenant d'un même enregistrement, car il s'agit de la stratégie permettant d'obtenir, expérimentalement, les plus faibles $\text{DER}_{\text{de collections}}$. Les deux stratégies de regroupement font l'objet d'une analyse détaillée en partie 5.3.

5.2. Perfectionnement des approches de regroupement

Nous profitons de ce premier chapitre sur les architectures de regroupement adaptées au traitement des collections pour présenter les modifications apportées aux méthodes de classification employées dans le cadre de la SRL, à savoir, le regroupement agglomératif hiérarchique (HAC) et le regroupement par Programmation Linéaire en Nombres Entiers (ILP). Ces améliorations ne sont pas spécifiques au traitement des collections, et sont d'ailleurs mises en œuvre dans notre système de SRL d'émissions. Nous présentons d'abord une reformulation de l'expression du problème ILP, visant à réduire la complexité des problèmes soumis à l'outil de résolution en limitant le nombre de variables et de contraintes. Cette nouvelle formulation allège les problèmes ILP soumis à l'outil de résolution. Nous proposons ensuite une configuration basée sur la modélisation i-vector et une évaluation PLDA. Nous présentons finalement une approche de classification en locuteur reposant sur la théorie des graphes. Cette méthode, appliquée en amont du procédé de regroupement, permet de déterminer des composantes connexes correspondant à des *sous-problèmes* de regroupement indépendants, et d'en résoudre la plupart de manière triviale.

5.2.1 Reformulation du problème de regroupement ILP

Nos travaux préliminaires nous ont permis de mettre en avant la viabilité de l'approche ILP dans le contexte du traitement des collections. Nous avons cependant constaté un inconvénient avec cette approche : la méthode *Branch & Bound* est un algorithme générique permettant de déterminer la solution optimale de problèmes d'optimisation discrets en énumérant systématiquement l'ensemble des solutions possibles (les bonnes comme les mauvaises). L'algorithme *Branch & Bound* ne peut donc pas être exécuté en temps polynomial, et lorsque le problème donné est trop volumineux (lorsque le nombre de classes, au départ, est trop conséquent) ou trop complexe

(lorsque les possibilités de regroupement sont trop nombreuses), le temps de calcul nécessaire à la résolution d'un problème peut devenir déraisonnable, et les fichiers d'entrée et de sortie du problème, trop lourds.

Cependant, une analyse préalable des problèmes peut permettre de rejeter un grand nombre des *mauvaises* solutions. Nous avons proposé à cet effet une reformulation du problème de regroupement ILP [Dupuy et al., 2014a] permettant d'éliminer la plupart des *mauvaises* solutions avant la soumission du problème à l'outil de résolution. Cette reformulation est sans conséquence sur la solution proposée, elle ne fait que réduire (drastiquement) le nombre de variables et de contraintes du problème. Dans la formulation originale du regroupement ILP, telle que présentée dans la partie 2.3.4 et rappelée ci-dessous, la contrainte 2.20e est la seule à faire intervenir la notion de score entre les classes :

$$Minimiser : \quad \sum_{k=1}^{N} x_{k,k} + \frac{1}{(S+1)} \sum_{k=1}^{N} \sum_{j=1}^{N} s(k,j) x_{k,j} \tag{2.20a}$$

$$Contraintes : \quad x_{k,j} \in \{0,1\} \qquad k \in C, j \in C \tag{2.20b}$$

$$\sum_{k=1}^{N} x_{k,j} = 1 \qquad j \in C \tag{2.20c}$$

$$x_{k,j} - x_{k,k} \leq 0 \qquad k \in C, j \in C \tag{2.20d}$$

$$s(k,j) x_{k,j} < \delta \qquad k \in C, j \in C \tag{2.20e}$$

Cette notion de score peut cependant être appliquée à chacune des autres contraintes. Le problème ILP peut alors être restreint aux seules variables binaires $x_{k,j}$ pour lesquelles $s(k,j) < \delta$, au lieu d'exprimer librement les contraintes 2.20b, 2.20c et 2.20d en fonction de toutes les valeurs de k et de j. Le score entre les modèles i-vector est nécessairement calculé avant la formulation du problème ILP. Nous avons donc proposé de reformuler la fonction objective à minimiser, ainsi que ses contraintes, en ne considérant que l'ensemble des valeurs possible de k pour lesquelles les scores entre les classes k et j sont inférieurs au score δ. Soit $C \in \{1 \dots N\}$, et $K_{j \in C} = \{k / s(k,j) < \delta\}$:

$$Minimiser : \quad \sum_{k \in C} x_{k,k} + \frac{1}{(S+1)} \sum_{j \in C} \sum_{k \in K_j} s(k,j) x_{k,j} \tag{5.1a}$$

$$Contraintes : \quad x_{k,j} \in \{0,1\} \qquad k \in K_j, j \in C \tag{5.1b}$$

$$\sum_{k \in K_j} x_{k,j} = 1 \qquad j \in C \tag{5.1c}$$

$$x_{k,j} - x_{k,k} \leq 0 \qquad k \in K_j, j \in C \tag{5.1d}$$

Comparé à la formulation originale, la contrainte 2.20e : $s(k,j)x_{k,j} < \delta \quad k \in C, j \in C$ est implicitement prise en compte dans les contraintes 5.1b, 5.1c et 5.1d en utilisant l'ensemble K_j à la place de C. S correspond à la somme de tous les scores inférieurs au seuil δ.

Étant donné une segmentation d'entrée composée de n classes, la formulation originale du regroupement ILP 2.20a générera n^2 variables binaires et $2n^2 - n$ contraintes. L'équation 2.20b ne génère aucune contrainte dans le problème, elle est implicitement générée par l'outil de résolution. L'équation 2.20c génère une contrainte pour chaque classe du problème, donc n contraintes. L'équation 2.20d génère $n-1$ contraintes par classe, *i.e.*, $n(n-1)$ puisqu'aucune contrainte n'est générée pour le cas où $k = j$. L'équation 2.20e génère quant à elle une contrainte pour chaque $s(k,j) < \delta$, ainsi qu'une autre contrainte pour chaque $s(k,j) > \delta$ (égale à $x_{k,j} = 0$). Cette dernière contrainte n'est pas exprimée dans la formulation du problème ILP, de la même manière qu'aucune contrainte n'est générée pour l'équation 2.20d lorsque $k = j$. Finalement, l'équation 2.20e génère $n(n-1)$ contraintes.

Dans notre reformulation du problème ILP 5.1a, le nombre de contraintes dépend du seuil δ. Plus la valeur de δ est élevée, plus le nombre de variables et de contraintes augmente jusqu'à tendre vers $n + n(n-1)$. L'équation 5.1b génère n contraintes, et l'équation 5.1c en génère $n(n-1)$. Quant à l'équation 5.1d, elle génère une contrainte pour chaque $s(k,j) > \delta$ lorsque $k \neq j$. Avec la reformulation du problème que nous avons introduit, le nombre de contraintes est égal au nombre de variables.

▷ Exemple de réduction du nombre de variables et de contraintes

Afin d'illustrer l'intérêt de ce travail, nous présentons dans le tableau 5.1 une comparaison portant sur le nombre de variables binaires et de contraintes entre la formulation originale et notre reformulation du problème. Les nombres de variables et de contraintes ont été déterminés par lecture des fichiers soumis à l'outil de résolution. Les données sur lesquelles porte cette comparaison correspondent aux émissions du corpus de test fourni durant le défi REPERE en janvier 2013.

En moyenne, sur les données du corpus de test REPERE de janvier 2013, notre reformulation du problème ILP permet de réduire le nombre de variables de 1744 à seulement 53, et le nombre de contraintes de 3449 à 53 également. Autrement dit, avec notre reformulation du problème ILP, l'algorithme *Branch & Bound* ne travaille qu'avec, en moyenne, 53 variables et 53 contraintes, contre 1744 et 3449 pour la formulation originale. En d'autres termes, la reformulation du regroupement ILP mène à une réduction du nombre de variables d'environ 97%, et une réduction

Émission	n^{bre} C.	Formulation originale (eq. 3.2a)		Reformulation (eq. 5.1a)	
		n^{bre} var.	n^{bre} contr.	n^{bre} var.	n^{bre} contr.
BFMStory 2012-01-10	68	4624	9180	84	84
BFMStory 2012-01-23	76	5776	11476	86	86
BFMStory 2012-02-14	67	4489	8911	87	87
BFMStory 2012-02-20	77	5929	11781	95	95
CultureEtVous 2012-01-13	13	169	325	17	17
CultureEtVous 2012-01-16	15	225	435	23	23
CultureEtVous 2012-01-17	15	225	435	15	15
CultureEtVous 2012-01-18	17	289	561	17	17
CultureEtVous 2012-01-19	16	256	496	18	18
CultureEtVous 2012-02-14	18	324	630	18	18
CultureEtVous 2012-02-15	21	441	861	23	23
CaVousRegarde 2011-12-20	39	1521	3003	69	69
CaVousRegarde 2012-01-19	39	1521	3003	73	73
CaVousRegarde 2012-01-25	43	1849	3655	79	79
EntreLesLignes 2011-12-16	41	1681	3321	77	77
EntreLesLignes 2012-01-27	35	1225	2415	41	41
EntreLesLignes 2012-05-11	45	2025	4005	71	71
LCPInfo13h30 2012-01-24	55	3025	5995	59	59
LCPInfo13h30 2012-01-25	51	2601	5151	69	69
LCPInfo13h30 2012-01-27	39	1521	3003	47	47
PileEtFace 2011-11-19	28	784	1540	52	52
PileEtFace 2011-12-01	30	900	1770	42	42
PileEtFace 2012-01-12	29	841	1653	41	41
PileEtFace 2012-01-19	42	1764	3486	72	72
PileEtFace 2012-01-26	38	1444	2850	70	70
TopQuestions 2012-01-25	33	1089	2145	41	41
TopQuestions 2012-02-14	26	676	1326	36	36
TopQuestions 2012-02-22	40	1600	3160	62	62
Minimum	-	169	325	15	15
Moyenne	-	1743.36	3449	53	53
Maximum	-	5929	11781	95	95

Table *5.1 – Nombre de variables et nombre de contraintes déterminé à partir des problèmes ILP soumis à l'outil de résolution, pour la formulation originale et notre reformulation du problème ILP, avec* $\delta = 105$ *(résultats par émissions.* n^{bre} *C. correspond au nombre de classes présentes dans les segmentations d'entrée).*

du nombre de contraintes d'environ 98% (en moyenne).

5.2.2 Comparaison de méthodes de classification

Nous avons commencé à travailler avec la modélisation i-vector dans le cadre du développement de la méthode de classification ILP. C'est en cherchant à perfection-

ner cette approche de regroupement que nous nous sommes intéressés aux techniques à l'état de l'art du domaine de la reconnaissance du locuteur. Sylvain Meignier, qui a évalué dans son HDR [Meignier, 2015] différentes approches de modélisation du locuteur, et différentes approches pour estimer la vraisemblance entre ces modèles, a montré que les approches i-vector et PLDA permettent d'obtenir les taux d'erreur les plus faibles pour la SRL d'émissions. Étant donné ses conclusions, nous proposons une configuration similaire, reposant sur des modèles i-vector de dimension 300, pour modéliser les classes de locuteurs, et sur l'opposé des scores PLDA pour estimer la similarité entre les classes. Nous proposons, dans les parties suivantes, une comparaison des méthodes de classification ILP (regroupement combinatoire en programmation linéaire en nombres entiers) et HAC (regroupement agglomératif hiérarchique) entre leurs configurations PLDA (nommées respectivement ILP_{PLDA} et HAC_{PLDA}), et leurs configurations alternatives à l'état de l'art avec la distance de Mahalanobis et le rapport de vraisemblance croisé (nommées respectivement ILP_{Maha} et HAC_{CLR}). Cette étude expérimentale, qui vise à démontrer l'intérêt de notre approche de classification orientée PLDA dans le cadre du regroupement en locuteurs pour les collections, est effectuée sur les sept collections du niveau *programme*, dont les caractéristiques sont rappelées dans le tableau 5.2.

Niveau	Collection		n^{bre} émi.	Durée		n^{bre} locuteurs	
				audio	UEM	total	récur.
Programme	BFMTV	**BFM Story** *(+ Ruth Elkrief)*	48	49:32	23:00	556	83
		Planète Showbiz *(+ Culture et Vous)*	160	38:27	5:00	771	75
	LCP	**Ça vous Regarde**	23	20:55	7:14	173	11
		Entre les Lignes	27	16:14	7:05	18	9
		LCP Info *(+ LCP Actu)*	48	20:34	11:14	317	96
		Pile et Face	33	19:44	6:11	46	15
		Top Questions	35	12:20	7:28	119	36

Table *5.2 – Collections d'émissions du niveau* programme.

Nous présentons dans un premier temps les résultats de la SRL d'émissions, dont l'union des segmentations de chaque émission de la collection sert de point de départ pour le regroupement global (*cf.* architecture illustrée en figure 5.1). Nous présentons ensuite une description de chaque configuration étudiée, puis nous discutons les résultats obtenus pour chacune des sept collections évaluées.

▷ SRL d'émissions

Le système de SRL d'émissions ayant été utilisé pour produire des segmentations locales aux enregistrements correspond à celui présenté dans la partie état de l'art (*cf.* chapitre 2) avec un regroupement ILP. Les classes déterminées durant l'étape de regroupement hiérarchique BIC ($\lambda = 3$), sujettes au regroupement ILP, sont modélisées par des modèles i-vector de dimension 300 extraits à l'aide d'un GMM-UBM constitué de 1024 composantes gaussiennes. La figure 5.2 présente les résultats d'une évaluation au niveau *émissions* pour les 7 collections d'émissions du niveau *Programme* (les segmentations produites pour chaque enregistrement d'une collection sont évaluées séparément).

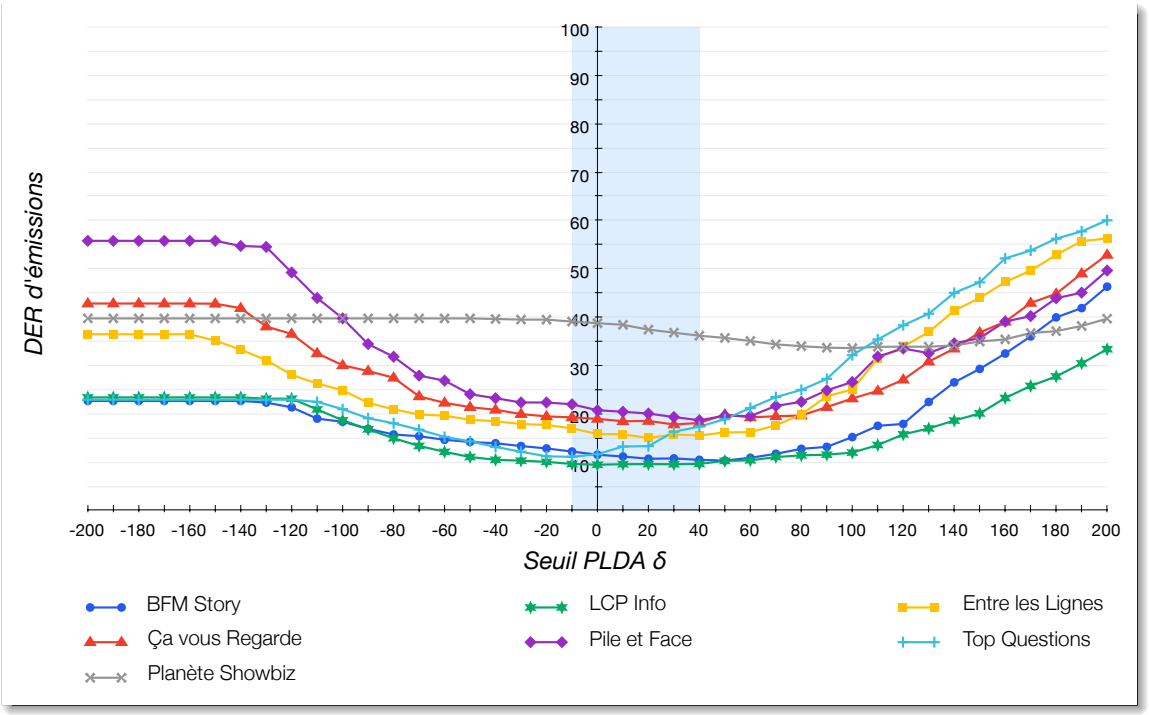

Figure *5.2* – **$DER_{d'émissions}$** *obtenus sur les 7 collections du niveau* Programme *avec le système de SRL d'émissions (regroupement ILP), pour différentes valeurs du seuil PLDA* δ.

Les $DER_{d'émissions}$ présentés ont été obtenus en faisant varier le seuil δ, qui correspond à la valeur du score PLDA à partir de laquelle un regroupement entre deux classes n'est plus toléré. En moyenne sur l'ensemble des enregistrements des sept collections étudiées, le meilleur taux d'erreur $DER_{d'émissions}$ (15,22%) est obtenu pour un seuil PLDA $\delta = 20$. L'évolution du $DER_{d'émissions}$ en fonction des valeurs du seuil δ suit une tendance sensiblement identique pour chaque collection étudiée, exception faite de la collection *Planète Showbiz*. Les résultats médiocres obtenus sur cette collection, quelque soit le seuil δ, s'expliquent principalement par le fait que l'émission *Planète Showbiz* représente un type de collections isolé (magazine *people*), et

le bruit inhérent à la nature des émissions est particulièrement varié : trois ou quatre locuteurs sont en général présents, l'ambiance est *relâchée* et spontanée, les rires sont francs, on s'amuse ... Les locuteurs ont tendance à commenter à tout va le sujet évoqué, sans se soucier de couper la parole aux autres. Il est également fréquent que des extraits musicaux, en rapport avec le sujet évoqué, couvrent en partie la voix des locuteurs. Le développement du système de SRL n'a pas été prévu pour prendre en compte les spécificités de ces enregistrements, qui s'apparentent plus à des enregistrements de réunions qu'à des enregistrements de journaux d'information. Avec la configuration utilisée, les classes issues du regroupement BIC contiennent souvent des segments de parole provenant de plusieurs locuteurs. Donc, d'une part, ces erreurs de regroupement ne sont pas réversibles, et d'autre part, les modèles i-vector extraits pour le regroupement ILP ne sont pas suffisamment discriminants, ce qui influence négativement la classification ILP. Un autre facteur d'influence est la durée évaluée par rapport à la durée totale des enregistrements. Ce rapport est d'environ 13% avec la collection *Planète Showbiz*, contre 45%, en moyenne, pour les autres collections étudiées. Or, les systèmes de SRL que nous présentons travaillent sur l'intégralité des fichiers audio, pas seulement sur les portions annotées à des fins d'évaluation. La portion évaluée des enregistrements n'est pas suffisamment représentative du volume total de la collection *Planète Showbiz*.

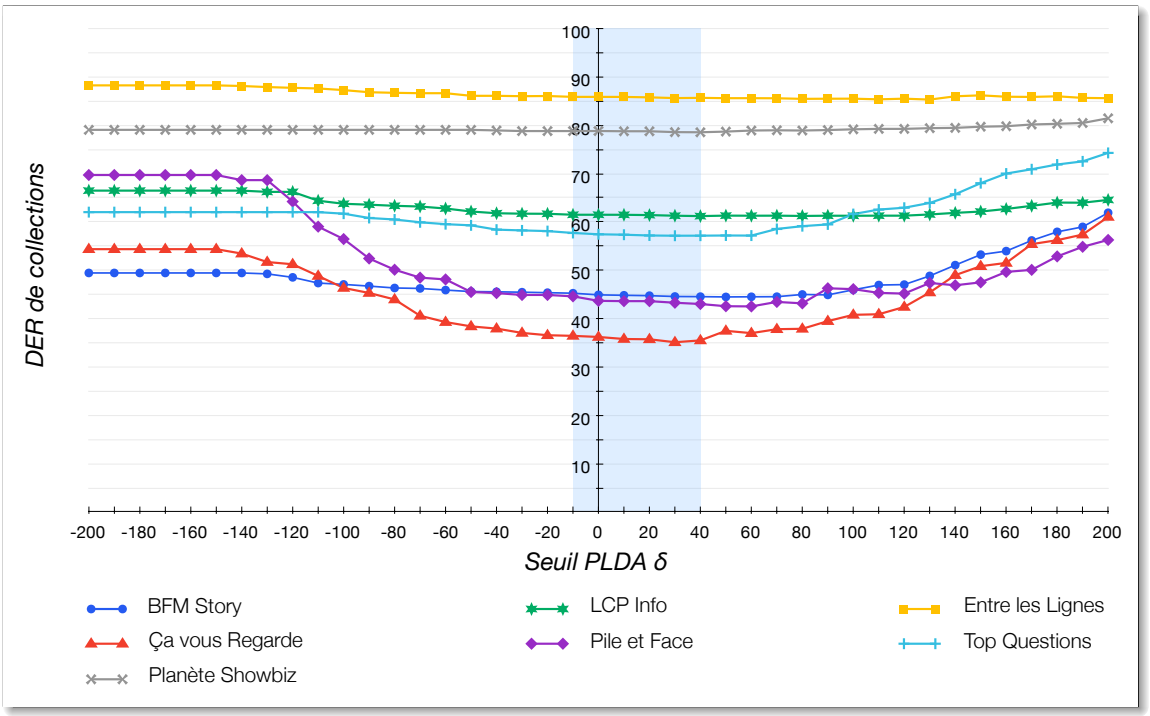

Figure 5.3 – **DER$_{\text{de collections}}$** *obtenus sur les 7 collections du niveau* Programme *avec le système de SRL d'émissions (regroupement ILP), pour différentes valeurs du seuil PLDA δ.*

À noter que le taux d'erreur DER$_{\text{d'émissions}}$ pour un seuil PLDA $\delta = 20$ descend

à 13,17% si la collection *Planète Showbiz* est exclue de l'évaluation. En figure 5.3, nous présentons les résultats d'une évaluation au niveau *collection*. Cette évaluation permet d'apprécier la difficulté de la SRL de collections, en révélant les taux d'erreur $\text{DER}_{\text{de collections}}$ avant que tout regroupement entre classes issues d'enregistrements différents ne soit effectué (les segmentations évaluées sont toujours celles produites par le système de SRL d'émissions). Au préalable, nous nous sommes assurés que les étiquettes attribuées aux classes de locuteurs soient bien dépendantes des enregistrements pour ne pas fausser les $\text{DER}_{\text{de collections}}$. Les $\text{DER}_{\text{de collections}}$ ainsi obtenus sont très élevés du fait de l'absence de regroupement inter-enregistrements, ils varient de 36% à 86% selon les collections lorsque le seuil PLDA δ est fixé à 20.

La difficulté de la SRL de collections se traduit donc par une brutale augmentation entre les $\text{DER}_{\text{d'émissions}}$ et $\text{DER}_{\text{de collections}}$ lorsque ces deux métriques sont déterminées sur les segmentations du niveau émission de l'architecture. Il convient cependant de relativiser quant aux $\text{DER}_{\text{de collections}}$ obtenus. Nous présentons à cet effet, dans le tableau 5.3, une comparaison entre les $\text{DER}_{\text{d'émissions}}$ et $\text{DER}_{\text{de collections}}$ obtenus à partir des segmentations d'émissions fournies par le système. Nous confrontons ces résultats aux $\text{DER}_{\text{de collections}}$ qu'il serait possible d'atteindre si les segmentations d'émissions avaient été sans erreurs. Nous avons pour cela réalisé une évaluation des collections à partir de segmentations de référence, en nous assurant au préalable que les étiquettes de locuteurs soient dépendantes des enregistrements, afin de simuler des segmentations parfaites au niveau *émission* de l'architecture seulement (nous simulons ainsi l'absence de regroupements entre les enregistrements).

Collection	**$\text{DER}_{\text{d'émissions}}$ δ = 20**	**$\text{DER}_{\text{de collections}}$ δ = 20**	**$\text{DER}_{\text{de collections}}$ références**	**$\text{DER}_{\text{de collections}}$ diff. sys - ref**
BFM Story	10,70%	44,75%	39,56%	5,19%
Planète Showbiz	37,41%	78,76%	60,52%	18,24%
Ça vous Regarde	18,51%	35,73%	21,76%	13,97%
Entre les Lignes	15,01%	85,77%	84,78%	0,99%
LCP Info	9,60%	61,42%	56,33%	5,09%
Pile et Face	20,05%	43,62%	28,00%	15,62%
Top Questions	13,31%	57,20%	50,71%	6,49%

Table *5.3 – Différence observée entre les* **$\text{DER}_{\text{d'émissions}}$** *et* **$\text{DER}_{\text{de collections}}$** *sur les segmentations d'émissions fournies par le système avec un seuil PLDA δ fixé à 20, pour sept collections étudiées.*

Les $\text{DER}_{\text{d'émissions}}$ de ces segmentations de référence ne sont pas présentés dans le tableau 5.3, leurs valeurs étant égaux 0,0%. L'écart entre les $\text{DER}_{\text{de collections}}$ des segmentations fournies par le système et les segmentations de référence dépend des collections, mais n'est pas aussi important que ce à quoi nous pouvions nous attendre (*cf.* dernière colonne du tableau 5.3). *Planète Showbiz*, *Ça vous Regarde* et *Pile et Face* sont les collections pour lesquelles le procédé de SRL de collections semble être

le plus difficile, compte tenu des segmentations d'émission fournies par le système : la différence entre les $DER_{\text{de collections}}$ des segmentations système et de référence est comprise entre 14% et 18% (il s'agit également des collections pour lesquelles les $DER_{\text{d'émissions}}$ sont les plus élevés). En revanche, les segmentations fournies par le système pour les autres collections ne détériorent pas de beaucoup le $DER_{\text{de collections}}$ par rapport à celui simulé avec les segmentations de référence, en particulier pour la collection *Entre les Lignes*, où la différence n'est que de 1% (il s'agit cependant de la collection la moins abondante en termes de locuteurs (18) et de locuteurs récurrents (9)).

▷ Approche de regroupement global par ILP

Configuration ILP_{Maha}

L'étude présentée par [Rouvier et Meignier, 2012] montre que l'approche de regroupement par ILP permet d'obtenir de meilleurs résultats que l'approche hiérarchique lorsque les classes de locuteurs sont représentées par des modèles i-vector. Nous avons suivi la même « recette », en proposant toutefois d'employer des modèles i-vector de dimension 300. Un modèle i-vector est extrait pour chacune des classes de locuteur issues de la segmentation concaténée. La paramétrisation correspond à 20 paramètres MFCC, leurs 20 coefficients différentiels Δ respectifs et 17 coefficients $\Delta\Delta$ (configuration standard de la suite d'outils pour la reconnaissance du locuteur *Alize* [Bonastre et al., 2008]). Le modèle du monde GMM-UBM, indépendant du genre et de la bande de fréquence est constitué de 1024 composantes gaussiennes, entraînées sur les données d'apprentissage des campagnes ESTER 1 & 2 et ETAPE [1] à l'aide de la suite *Alize*. La mesure de vraisemblance utilisée pour estimer la similarité entre les modèles i-vector est la distance de Mahalanobis. La normalisation des modèles i-vector est effectuée par 5 itérations de l'algorithme EFR (EigenFactors Radial normalization). Le programme d'optimisation linéaire utilisé pour résoudre le problème ILP, à l'aide de l'algorithme *Branch & Bound*, est l'outil de résolution *glpsol*, distribué librement dans la suite GPL *GNU GLPK*.

Les résultats en termes de $DER_{\text{de collections}}$ et $DER_{\text{d'émissions}}$, obtenus par l'approche de regroupement global ILP_{Maha} sur les sept collections du niveau *programme*, sont présentés dans les figures 5.4 et 5.5.

Les taux d'erreur DER dépendent du seuil δ, qui correspondant à la valeur de la distance de Mahanalobis pour laquelle deux classes (modèles i-vector) ne peuvent

1. Les données d'apprentissage du corpus ETAPE n'ont pas été incluses dans nos collections (*cf.* partie 4.3).

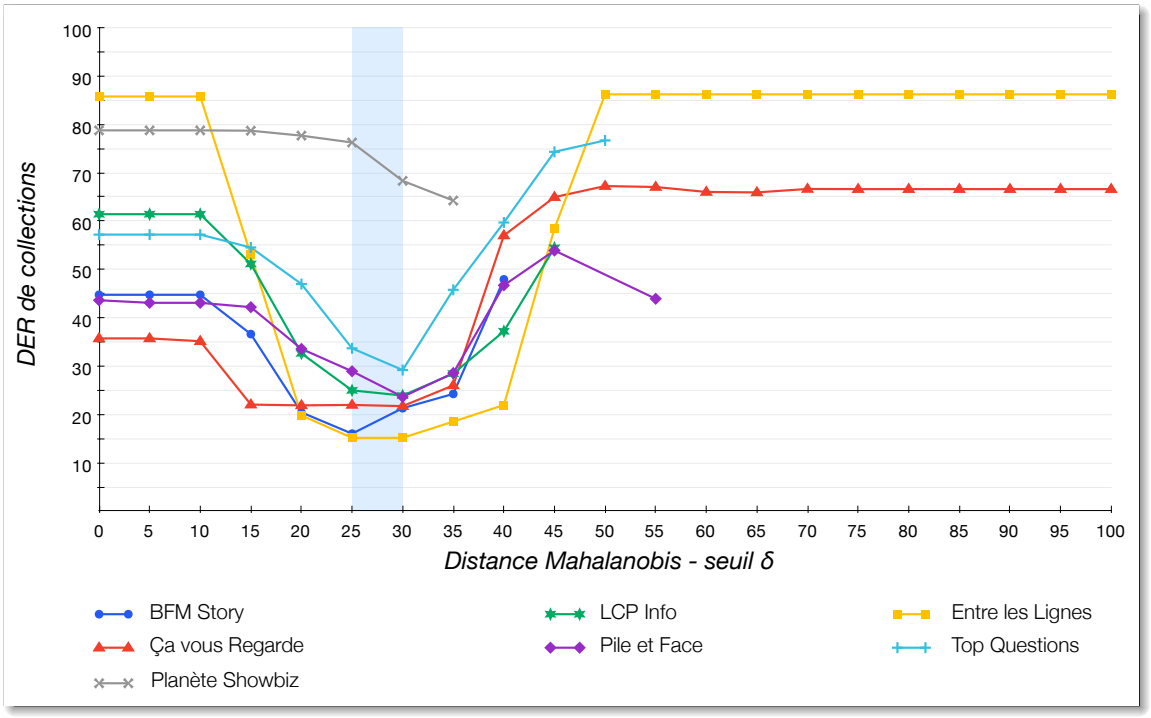

Figure 5.4 – **DER**$_{\text{de collections}}$ *obtenus sur les 7 collections du niveau* Programme *avec un regroupement* global *ILP, pour différentes valeurs de la distance de Mahalanobis* δ.

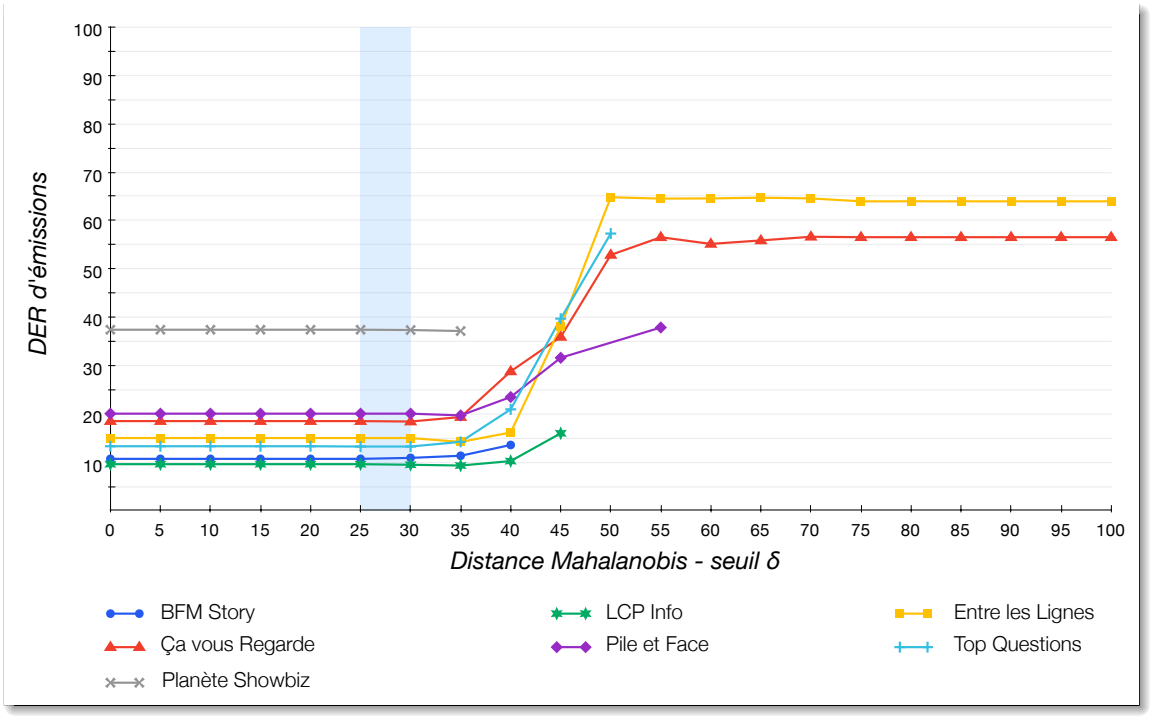

Figure 5.5 – **DER**$_{\text{d'émissions}}$ *obtenus sur les 7 collections du niveau* Programme *avec un regroupement* global *ILP, pour différentes valeurs de la distance de Mahalanobis* δ.

pas être regroupées. Plus la valeur de δ augmente, plus le nombre de variables et de contraintes définissant les problèmes ILP augmente. À partir d'un certain seuil δ

(dénommé par la suite « δ plafond »), les problèmes soumis à l'outil de résolution sont trop complexes[2] pour être résolus dans un laps de temps raisonnable. Nous considérons que la durée du regroupement pour une valeur donnée de δ ne doit pas excéder la durée de la plus grande collection étudiée. La valeur *plafond* du seuil δ dépend de la collection traitée et du nombre de classes impliquées dans le problème de regroupement, c'est pourquoi certains résultats n'apparaissent pas sur les graphiques présentés dans les figures 5.5 et 5.4. En moyenne sans tenir compte de la collection *Planète Showbiz*, qui est trop atypique, le plus faible taux $\text{DER}_{\text{de collections}}$ (21,47%) est obtenu pour un seuil δ égal à 25. L'évolution des $\text{DER}_{\text{de collections}}$ est similaire entre les collections, exception faite de *Planète Showbiz*, cependant la zone de stabilité est très réduite (aire rectangulaire sur les graphiques). Les $\text{DER}_{\text{d'émissions}}$, présentés en figure 5.5, sont stables jusqu'à $\delta = 30$ et quasiment identiques à ceux obtenus avant par le système de SRL d'émissions, avant le regroupement global. En revanche, passé la valeur *plafond* du seuil δ, les résultats en termes de $\text{DER}_{\text{d'émissions}}$ se dégradent considérablement, et rapidement. Le tableau 5.4 présente pour chaque collection le nombre de classes impliquées dans le problème de regroupement (n^{bre} classes initiales) et la valeur *plafond* du seuil δ.

Collection	n^{bre} enr.	n^{bre} classes initiales	Seuil δ plafond	n^{bre} scores $> \delta$ plafond
Entre les Lignes	27	732	> 100	0
Ça vous Regarde	23	814	> 100	0
Pile et Face	33	868	60	53550
Top Questions	35	1000	55	241730
LCP Info	48	1812	50	1724702
BFM Story	48	2845	45	6798766
Planète Showbiz	160	4224	40	17407794

Table *5.4 – Corrélation entre le nombre de classes d'une collection et la valeur* plafond *du seuil δ pour le regroupement global de type ILP_{Maha}. La dernière colonne indique le nombre de scores entre deux classes pour le seuil δ plafond observé (attention, les matrices de scores sont symétriques).*

Dans ce tableau, les collections sont ordonnées en fonction du nombre de classes impliquées dans les problèmes de regroupement, permettant ainsi de constater que plus le nombre de classes augmente, plus la valeur *plafond* du seuil δ, rendant impossible l'obtention d'une solution en un laps de temps raisonnable, tend vers 0 (et plus le nombre de scores supérieurs au seuil δ plafond augmente). Seules *Ça vous Regarde* et *Entre les Lignes*, qui correspondent aux collections les moins fournies en classes (respectivement 814 et 732), ne sont pas contraintes par une valeur *plafond* du seuil δ (du moins, cette valeur n'est pas atteinte pour $\delta = 100$).

2. Le nombre de classes impliquées dans le problème est trop conséquent, et les possibilités de regroupements sont trop nombreuses

Configuration ILP$_{\text{PLDA}}$

Cette configuration est quasiment identique à la précédente. Elle ne diffère que par les scores, qui correspondent aux opposés des scores PLDA entre les classes, et par l'étape de normalisation des modèles i-vector, qui est réalisée au moyen de l'algorithme SNN (Spherical Nuisance Normalization) avec une itération. Cette configuration ne permet pas de résoudre le problème de seuil δ *plafond* observé avec la distance de Mahalanobis, elle permet cependant de le repousser. En moyenne sur l'ensemble des collections (*Planète Showbiz* mise à part), la plage de seuil δ permettant d'approcher le meilleur taux d'erreur DER$_{\text{de collections}}$ est localisée entre -60 et 0, le meilleur DER$_{\text{de collections}}$ (19,43%) étant obtenu pour un seuil δ = -30 (*cf.* figure 5.6).

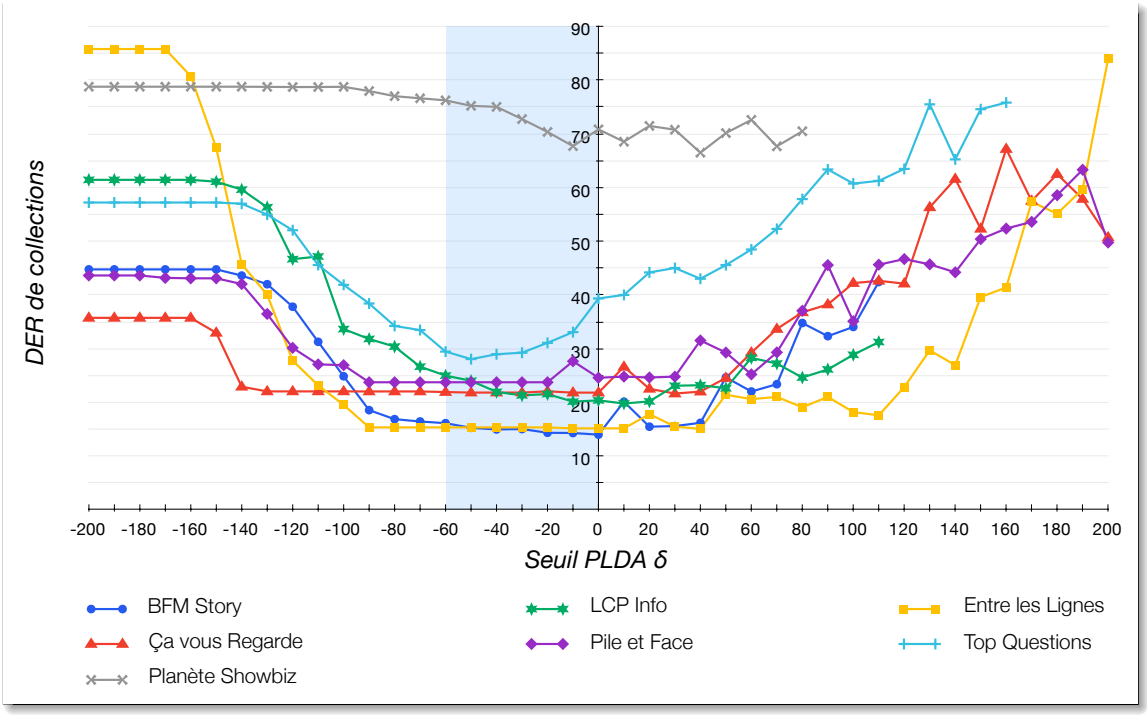

Figure *5.6* – **DER$_{\text{de collections}}$** *obtenus sur les 7 collections du niveau* Programme *avec un regroupement* global *ILP, pour différentes valeurs du score PLDA* δ.

La configuration PLDA permet d'atteindre un taux d'erreur DER$_{\text{de collections}}$ inférieur à celui de la configuration état de l'art reposant sur la distance de Mahalanobis (-2,04% en moyenne). Les résultats en termes de DER$_{\text{d'émissions}}$, présentés en figure 5.7, sont semblable à ceux obtenus au niveau *émission* de l'architecture, et stables jusqu'à seuil δ = 80. Contrairement à l'approche état de l'art où ces mêmes taux d'erreur se dégradent rapidement lorsque le seuil δ est à peine plus élevé que celui donnant les meilleurs résultats en termes de DER$_{\text{de collections}}$, l'approche PLDA offre plus de souplesse quant au choix du seuil δ, notamment par la stabilité des résultats

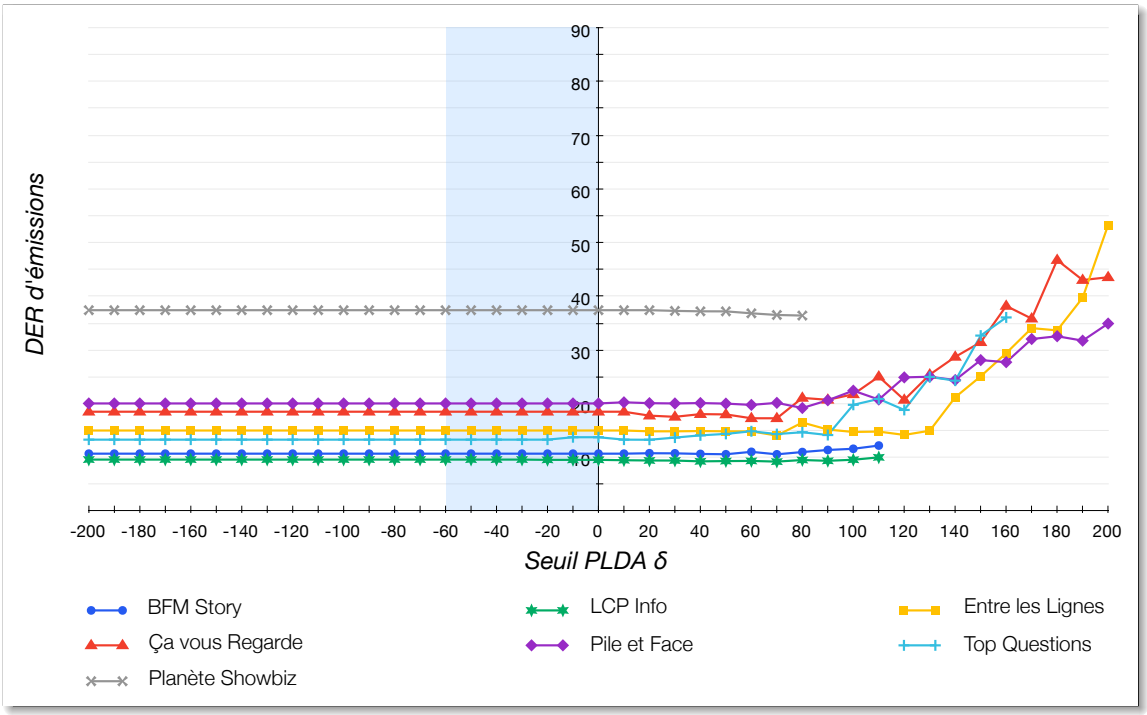

Figure 5.7 – **DER$_{\text{d'émissions}}$** *obtenus sur les 7 collections du niveau* Programme *avec un regroupement* global *ILP, pour différentes valeurs du score PLDA* δ.

jusqu'à des valeurs élevées de δ. Le tableau 5.5 présente la corrélation entre les seuils δ *plafond*, à partir desquels la résolution du problème de regroupement ne semble pas aboutir, et le nombre de classes impliquées dans les problèmes de regroupement, pour chacune des collections étudiées.

Collection	n$^{\text{bre}}$ enr.	n$^{\text{bre}}$ classes initiales	Seuil δ plafond	n$^{\text{bre}}$ scores $> \delta$ plafond
Entre les Lignes	27	732	> 200	419174
Ça vous Regarde	23	814	> 200	531320
Pile et Face	33	868	> 200	578316
Top Questions	35	1000	170	852868
LCP Info	48	1812	120	3192176
BFM Story	48	2845	120	7914616
Planète Showbiz	160	4224	90	17661762

Table 5.5 – *Corrélation entre le nombre de classes d'une collection et la valeur* plafond *du seuil* δ *pour le regroupement global de type* ILP_{PLDA}. *La dernière colonne indique le nombre de scores entre deux classes pour le seuil* δ plafond *observé (attention, les matrices de scores sont symétriques).*

Le seuil δ *plafond* pour la collection la plus fournie en classes (Planète Showbiz) est de 90, donc largement supérieur à la valeur optimale de δ qui est de -30. Au regard des observations effectuées par comparaison avec la configuration à l'état de l'art, la configuration PLDA apparaît plus adaptée pour traiter des collections volumineuses par regroupement global ILP. En revanche, le problème de limitation (valeur *plafond*

de δ) lié à la quantité de classes candidates pour le regroupement persiste, et le regroupement ILP_{PLDA} risque d'échouer pour des collections volumineuses.

▷ Approche de regroupement global par HAC

Configuration HAC_{CLR}

Cette première version du regroupement agglomératif hiérarchique correspond à la configuration état de l'art ayant permis d'obtenir les premières places lors des campagnes d'évaluation ESTER 1 & 2 (*cf.* partie 2.3.3). Les classes de locuteur sont représentées par des modèles GMM par adaptation MAP d'un GMM-UBM. Ce modèle du monde, indépendant du genre et de la bande de fréquence, est composé de 512 composantes gaussiennes et entraîné sur les données d'apprentissage distribuées durant la campagne d'évaluation française ESTER 1. La paramétrisation acoustique correspond à 12 paramètres MFCC, dont l'énergie, complétés par leurs coefficients différentiels Δ respectifs. Ces paramètres sont normalisés par les méthodes de *features warping* et CMS.

Afin de gagner du temps sur l'estimation des nouveaux modèles GMM résultant de la fusion de deux classes, lors du regroupement hiérarchique, nous ne réalisons qu'une seule itération de l'algorithme MAP : les nouveaux modèles GMM sont obtenus par fusion des accumulateurs statistiques des modèles GMM des deux classes fusionnées (méthode évoquée par [Leeuwen, 2010] (*cf.* partie 3.1)). La mesure utilisée pour estimer la vraisemblance entre les modèles GMM est le rapport de vraisemblance croisé (CLR). Afin d'accélérer la phase de calcul des mesures CLR entre le nouveau modèle GMM (résultant d'une fusion), et les autres, on ne considère que les 5 meilleures composantes gaussiennes [Reynolds et al., 2000a].

Les modifications apportées de manière réduire la durée du regroupement sont essentielles dans le contexte du traitement des collections. Certes, les résultats en termes de DER ne sont pas aussi bons que ce qu'ils pourraient être si l'adaptation MAP avait été réalisée avec plusieurs itérations au lieu d'une, et si la mesure de vraisemblance avait été calculée sur plus de gaussiennes. Malgré ce compromis visant à réduire la complexité du regroupement global, la méthode HAC_{CLR} reste difficilement exploitable étant donné la contrainte que nous nous sommes imposée quant à la durée de la classification (la durée du regroupement global ne doit pas excéder la durée de la plus volumineuse des collections étudiées, c'est-à-dire, environ 178 heures).

Les figures 5.8 et 5.9 présentent les taux d'erreur $\text{DER}_{\text{de collections}}$ et $\text{DER}_{\text{d'émissions}}$ obtenus sur les sept collections du niveau *programme*, en faisant varier la valeur du

seuil de décision α, qui correspond à la valeur à partir de laquelle la similarité entre les classes n'est plus suffisamment élevée pour autoriser un regroupement.

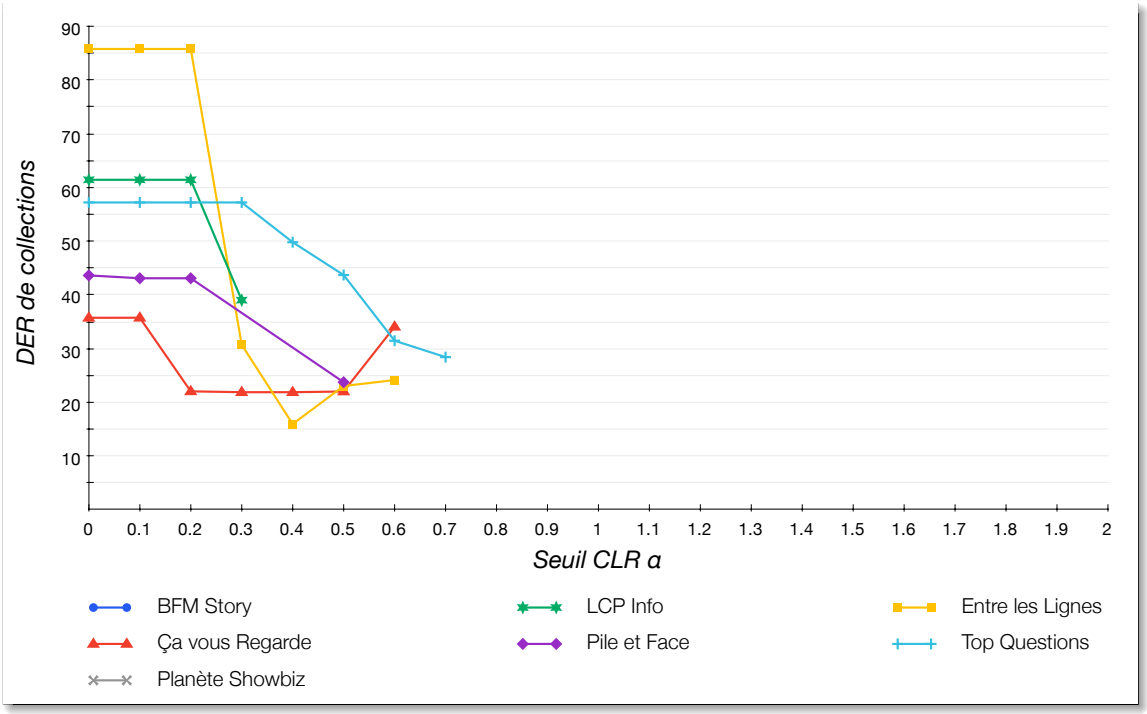

Figure 5.8 – **DER**$_{\text{de collections}}$ *obtenus sur les 7 collections du niveau* Programme *avec un regroupement* global *HAC, pour différentes valeurs du seuil CLR* α.

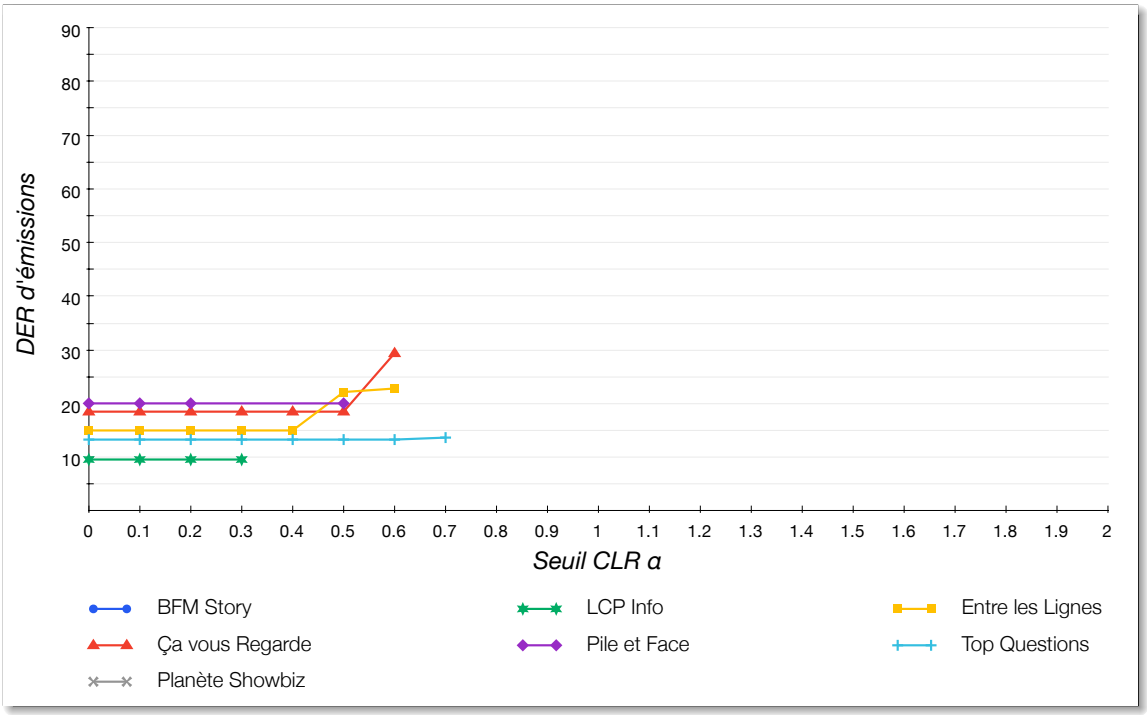

Figure 5.9 – **DER**$_{\text{d'émissions}}$ *obtenus sur les 7 collections du niveau* Programme *avec un regroupement* global *HAC, pour différentes valeurs du seuil CLR* α.

Des expériences antérieures sur des collections différentes, moins volumineuses que celles étudiées dans le cadre de ce manuscrit de thèse, nous ont permis d'estimer un seuil α optimal proche de 1.0 (*cf.* annexe C). Or, les collections du niveau *programme* sont trop volumineuses en termes de classes pour qu'un regroupement HAC_{CLR} avec un tel seuil α ne soit réalisé dans un laps de temps raisonnable (la durée du regroupement pour une valeur seuil donnée ne doit pas excéder la durée réelle de la plus grande des collections que nous avons proposées). Ce seuil α optimal reste néanmoins hypothétique : les collections étudiées dans le cadre de cette thèse sont constituées d'enregistrement d'émissions télévisuelles, et le seuil α optimal a été déterminé sur des collections d'enregistrements exclusivement radiophoniques avec une méthode classique pour la SRL d'émissions.

L'approche de regroupement HAC_{CLR} n'aboutit à aucun résultat, étant donné la contrainte de temps que nous nous sommes imposé, sur les deux collections les plus volumineuses du niveau *Programme* (*Planète Showbiz* et *BFM Story*), et ce même pour la plus faible valeur α testée. Il est difficile de discuter les résultats obtenus compte tenu de ce problème de faisabilité. Nous pouvons toutefois constater avec les résultats obtenus sur la collection *Entre les Lignes* que la métrique CLR permet d'atteindre des résultats similaires à ceux présentés pour les approches déjà étudiées, avec un $\text{DER}_{\text{de collections}}$ minimal de 15,95% pour un seuil α égal à 0.4.

La méthode de classification agglomérative hiérarchique est souvent utilisée dans l'état de l'art [Chen et Gopalakrishnan, 1998; Gish et al., 1991; Reynolds et al., 1998; Siegler et al., 1997; Siu et al., 1992; Solomonoff et al., 1998], mais elle n'est pas adaptée au traitement des collections volumineuses (le temps de calcul est trop long). La complexité de la classification HAC porte sur le nombre de trames (des vraisemblances sont recalculées après chaque regroupement), alors qu'elle porte sur le nombre de classes initiales pour la méthode de classification ILP.

Configuration HAC_{PLDA}

Dans cette deuxième configuration pour le regroupement HAC, la modélisation du locuteur est identique à celle présentée pour la méthode ILP_{PLDA}. Pour rappel, la paramétrisation acoustique repose sur 20 paramètres MFCC, agrémentés de leurs 20 coefficients Δ et de 17 coefficients $\Delta\Delta$. Des modèles i-vector de dimension 300 sont extraits pour chacune des classes de la segmentation initiale, correspondant à la réunion des segmentations obtenues au niveau émission de notre architecture. La mesure permettant d'estimer la similarité entre les classes correspond à une évaluation PLDA (les modèles i-vector sont normalisés par SNN *(Spherical Nuisance Normalization)*). Le regroupement HAC ne recalcule pas les modèles, le critère de

liaison utilisé correspond à celui du saut maximum (la *distance* entre deux classes correspond à la plus grande des distances entre les éléments des deux classes).

La figure 5.10 présente les taux d'erreur $DER_{\text{de collections}}$ obtenus par regroupement global HAC_{PLDA} sur les sept collections étudiées. Les valeurs testées du seuil PLDA δ s'échelonnent toujours de -200 à 200.

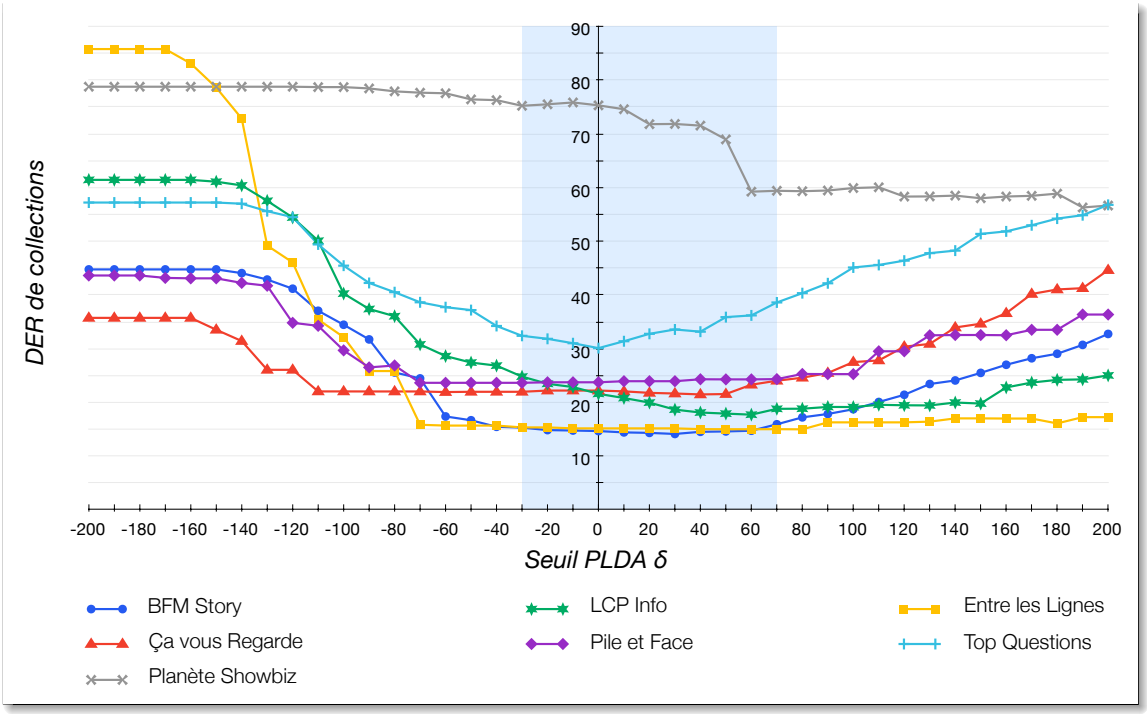

Figure *5.10* – **$DER_{\text{de collections}}$** *obtenus sur les 7 collections du niveau* Programme *avec un regroupement* global *HAC, pour différentes valeurs du seuil PLDA* δ.

La modélisation i-vector et l'approche d'évaluation PLDA permettent d'employer efficacement le regroupement hiérarchique dans le contexte de traitement des collections, contrairement à l'approche HAC_{CLR} où aucun résultat intéressant n'a pu être observé pour un temps de traitement raisonnable : la règle du saut maximum permet d'éviter le calcul de nouveaux scores et l'extraction de nouveaux modèles i-vectors. L'évolution des taux d'erreur en fonction de la valeur du seuil δ est relativement stable d'une collection à l'autre, exception faite de la collection *Planète Showbiz* pour laquelle les résultats restent décevants quelque soit le seuil δ. Cela n'a cependant rien de surprenant étant donné le caractère atypique de cette collection et les taux $DER_{\text{d'émissions}}$ élevés obtenus durant l'étape de SRL d'émissions. En moyenne sans tenir compte de la collection *Planète Showbiz*, les taux d'erreur $DER_{\text{de collections}}$ sont inférieurs ou égaux à 20% sur la plage de seuil s'étalant de -30 à 70 (aire bleu ciel sur les graphiques), avec un minimum égal à 19,08% pour le seuil $\delta = 30$. Contrairement à la configuration HAC_{CLR}, aucune valeur *plafond* du seuil δ n'est constatée dans la plage des valeurs évaluées.

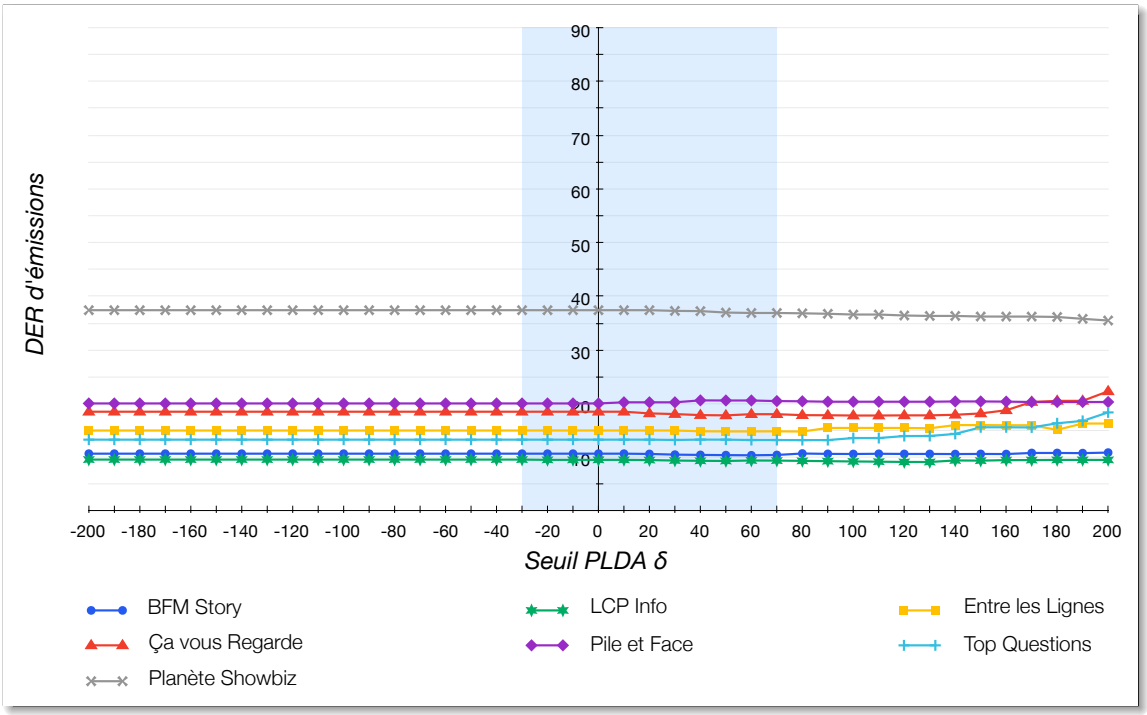

Figure *5.11* – **$DER_{d'émissions}$** *obtenus sur les 7 collections du niveau* Programme *avec un regroupement* global *HAC, pour différentes valeurs du seuil PLDA* δ.

Les $DER_{d'émissions}$ correspondants, présentés en figure 5.11, sont presque identiques à ceux obtenus par le système de SRL d'émission (avant l'étape de regroupement global). Ces $DER_{d'émissions}$ sont également très stables, quelle que soit la valeur du seuil PLDA, en particulier pour les valeurs de δ inférieures à 80. Ces deux observations, qui ont déjà été constatées lors de la présentation de l'approche de regroupement ILP_{PLDA}, peuvent là encore s'expliquer par le pouvoir discriminant des scores PLDA. En effet, les segmentations établies au niveau *émission* de l'architecture semblent être préservées par l'approche de regroupement global, bien que la liberté de regrouper des classes issues d'une même émission lui soit laissée.

▷ Comparaison et discussion

Nous proposons dans cette partie de comparer les quatre configurations de regroupement global, présentées et évaluées indépendamment dans les parties précédentes, dont les caractéristiques sont résumées dans le tableau 5.6.

Les résultats obtenus sur les collections du niveau *programme*, sans tenir compte de la collection *Planète Showbiz*, par les approches ILP_{Maha}, ILP_{PLDA} et HAC_{PLDA} sont présentés dans le tableau 5.7 pour les $DER_{de\ collections}$ et dans le tableau 5.8 pour les $DER_{d'émissions}$ (les résultats obtenus avec l'approche HAC_{CLR} ne sont pas suffisamment significatifs pour être évoqués). Les taux d'erreur présentés correspondent aux

	HAC_{CLR}	ILP_{Maha}	HAC_{PLDA}	ILP_{PLDA}
Paramétrisation	12 MFCC + 12Δ	20 MFCC + 20Δ + 17 ΔΔ		
GMM-UBM	512 gaussiennes	1024 gaussiennes		
Modélisation	GMM	i-vector (dimension 300)		
Normalisation	MVN + Feature Warping	EFR (5 itérations)	SNN (1 itération)	
Distance	CLR	Mahalanobis	PLDA (150 locuteurs, 40 sessions)	

Table *5.6 – Résumé des configurations évaluées.*

$\text{DER}_{\text{de collections}}$ obtenus pour le seuil δ minimisant le $\text{DER}_{\text{de collections}}$ moyen sur ces six collections. Attention, nous rappelons que les moyennes sont calculées en pondérant, pour chaque collection, les scores obtenus par la durée effectivement évaluée. Les résultats présentés entre parenthèses correspondent aux meilleurs $\text{DER}_{\text{de collections}}$ obtenus par collection et par configuration, tous seuils δ confondus. La colonne « Point de départ » présente les taux d'erreur DER déterminés sur les segmentations produites au niveau *émission* de l'architecture, c'est-à-dire, en amont du procédé de regroupement global. Ces valeurs permettent d'illustrer la difficulté de la tâche de regroupement global.

Configuration	**Point de départ**	$\textbf{ILP}_{\text{Maha}}$	$\textbf{ILP}_{\text{PLDA}}$	$\textbf{HAC}_{\text{PLDA}}$
Seuil δ	(20)	25	-30	30
BFM Story	44,75%	16,03% (16,03%)	15,03% (13,99%)	14,12% (14,12%)
Ça vous Regarde	35,73%	21,97% (21,72%)	21,82% (21,64%)	21,65% (21,48%)
Entre les Lignes	85,77%	15,16% (15,16%)	15,34% (15,06%)	15,16% (14,98%)
LCP Info	61,42%	24,98% (23,91%)	21,30% (19,78%)	18,65% (17,71%)
Pile et Face	43,62%	28,93% (23,64%)	23,74% (23,74%)	23,94% (23,64%)
Top Questions	57,20%	33,68% (29,19%)	29,24% (28,05%)	33,57% (30,10%)
Moyenne	**52,97%**	**21,47%**	**19,43%**	**19,08%**

Table *5.7 –* $\textbf{DER}_{\textbf{de collections}}$ *obtenus pour chaque collection étant donné un seuil δ optimal moyen. Les résultats entre parenthèses correspondent aux meilleurs* $\text{DER}_{\text{de collections}}$ *obtenus sur les collections tous seuils confondus.*

En termes de $\text{DER}_{\text{de collections}}$, les approches de regroupement utilisant les scores PLDA se montrent plus performantes que leurs alternatives, en particulier pour la méthode de regroupement HAC_{CLR} qui n'a pas permis d'aboutir à des résultats acceptables avec les scores CLR. La méthode de regroupement ILP_{PLDA} permet d'atteindre des taux d'erreur $\text{DER}_{\text{de collections}}$ inférieurs de 2% en moyenne à ceux obtenus par l'approche d'évaluation *Mahalanobis*. Les résultats obtenus avec la méthode de regroupement HAC_{PLDA} sont légèrement meilleurs, en moyenne, que ceux obtenus avec la méthode ILP_{PLDA}, avec un gain absolu de 0,35%. Ces résultats moyens sont toutefois à nuancer, car le comportement des deux méthodes de regroupement basées

sur l'évaluation PLDA n'est pas identique d'une collection à l'autre. Sur la collection *Top Question*, la méthode de regroupement ILP permet d'atteindre un $DER_{\text{de collections}}$ moyen inférieur de 4,33% en absolu à celui obtenu avec le regroupement HAC, et la constatation inverse peut être faite sur la collection *LCP Info* avec un écart absolu de 2,65% en faveur du regroupement HAC. Il est également bon de rappeler que les scores moyens n'ont été calculés que sur 6 collections, les résultats sur *Planète Showbiz* étant trop éloignés.

Configuration	**Point de départ**	**ILP_{Maha}**	**ILP_{PLDA}**	**HAC_{PLDA}**
Seuil δ	(20)	25	-30	30
BFM Story	10,70%	10,70% (10,70%)	10,70% (10,59%)	10,52% (10,40%)
Ça vous Regarde	18,51%	18,51% (18,42%)	18,51% (17,29%)	18,07% (17,80%)
Entre les Lignes	15,01%	15,01% (14,20%)	15,01% (14,07%)	15,01% (14,82%)
LCP Info	9,60%	9,60% (9,32%)	9,60% (9,27%)	9,42% (9,15%)
Pile et Face	20,05%	20,05% (19,69%)	20,05% (19,20%)	20,26% (20,05%)
Top Questions	13,31%	13,25% (13,25%)	13,31% (13,29%)	13,25% (13,23%)
Moyenne	**13,17%**	**13,16%**	**13,17%**	**13,03%**

Table *5.8* – **$DER_{\text{d'émissions}}$** *obtenus pour chaque collection étant donné un seuil δ optimal moyen. Les résultats entre parenthèses correspondent aux meilleurs $DER_{\text{d'émissions}}$ obtenus sur les collections tous seuils confondus.*

Les taux d'erreur $DER_{\text{d'émissions}}$, déterminés sur les segmentations obtenues au niveau *collection*, sont dans l'ensemble très proches de ceux obtenus au niveau *émission*. Les écarts observés, s'il y en a, sont minimes et positifs, ce qui conforte le choix de la stratégie de regroupement appliquée, dans laquelle la liberté est laissée au système de regrouper des classes provenant d'un même enregistrement (ce point sera discuté plus en détail dans la partie 5.3). On observe en particulier un gain notable pour la méthode de regroupement HAC_{PLDA}, permettant de réduire le $DER_{\text{d'émissions}}$ moyen de 0,14% en absolu par rapport à celui obtenu au niveau *émission* (colonne « Point de départ »).

Nous présentons en figure 5.12 un histogramme représentant la distribution des scores PLDA. Ces scores ont été recueillis à partir des matrices carrées recensant les scores entre chaque paire de modèles de locuteur, pour chacune des sept collections d'émissions de niveau *Programme* (en incluant donc les scores de la collection *Planète Showbiz*).

Contrairement à ce que cette figure laisse penser, la distribution des scores PLDA est bimodale. Seule une modalité est observable sur la figure 5.12, du fait de l'échelle des fréquences dont la valeur maximale est très élevée (nous sommes en présence de 32171389 scores, les matrices de scores étant symétriques). Il s'agit des scores correspondant au cas ou les deux locuteurs sont différents, la valeur maximum étant

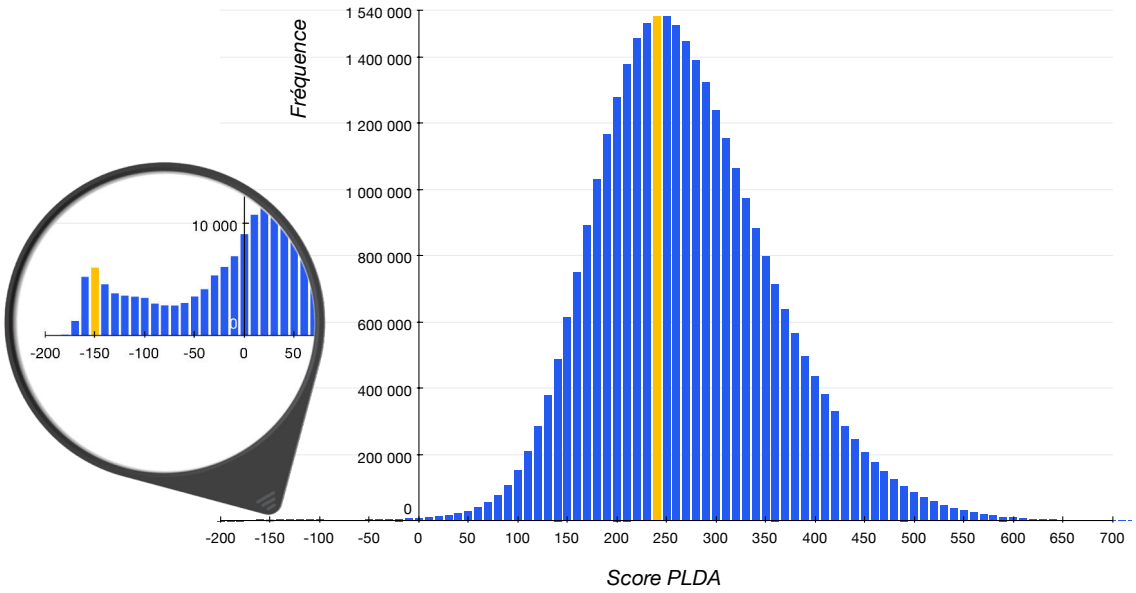

Figure *5.12 – Distribution des scores PLDA.*

proche de 240. Nous présentons sur la gauche de cette figure un « agrandissement » (la distribution est la même, seule l'échelle des fréquences a été modifiée), afin d'observer clairement les deux modalités dont les valeurs centrales sont représentées par des barres de couleur claire. On observe ainsi que la valeur du score PLDA calculé entre deux modèles de locuteurs censés représenter un même locuteur est proche de -150.

Bilan

Les deux méthodes de classification HAC et ILP ont été adaptées avec succès à la SRL de collections. La configuration la plus efficace en termes de résultats et temps de calcul repose sur une modélisation i-vector et sur des scores PLDA. Choisir entre les méthodes HAC et ILP est difficile, les résultats en termes de $\text{DER}_{\text{de collections}}$ et $\text{DER}_{\text{d'émissions}}$ sont très proches. Le choix de la méthode de classification devrait être effectué en fonction des contraintes techniques et algorithmiques à respecter : l'algorithme de classification HAC permet de déterminer une solution en temps polynomial, contrairement à l'algorithme de classification ILP qui risque de ne pas aboutir à une solution selon le problème traité.

5.2.3 Théorie des graphes et regroupement en locuteurs

La reformulation du problème ILP, présentée dans la partie précédente, a permis une diminution conséquente de la complexité des problèmes ILP à résoudre par l'algorithme *Branch & Bound*. L'idée derrière cette reformulation peut être généralisée : qu'il s'agisse de la méthode de classification HAC ou ILP, les classes candidates au regroupement sont modélisées **avant** l'exécution de l'algorithme de regroupement. Il en va de même pour les mesures de vraisemblance entre ces classes candidates, qui sont estimées avant le procédé de regroupement, quels que soient la mesure de vraisemblance et l'algorithme de classification choisis. Enfin, le seuil correspondant à la valeur à partir de laquelle un regroupement est proscrit est également établi à l'avance.

Nous avons proposé, étant donné ces constatations, une approche reposant sur la théorie des graphes pour simplifier le problème de classification. Ce travail qui, à l'origine, avait été pensé pour simplifier l'approche de classification ILP, peut cependant être appliqué à d'autres algorithmes de classification, dont l'approche HAC. Dans les sections suivantes, nous présentons notre approche de simplification telle que nous l'avons pensée pour la classification ILP. Nous présentons ensuite une évaluation avec les méthodes de classification ILP et HAC.

Tout comme la reformulation du problème ILP que nous avons présenté en partie 5.2.1, l'approche de simplification que nous proposons ici n'a pas d'incidence sur les classifications produites, les résultats en termes de DER restent les mêmes que ceux obtenus par l'approche ILP si cette simplification n'est pas mise en œuvre. La simplification que nous proposons est effectuée en temps polynomial, et appliquée en amont de l'algorithme de regroupement : elle permet de décomposer un problème de regroupement en sous-problèmes indépendants, dont la plupart sont triviaux à résoudre. Seuls les quelques sous-problèmes dits « complexes » doivent alors être résolus par l'algorithme de classification ILP. Cette approche permet donc d'éviter la recherche des regroupements triviaux par l'algorithme non-polynomial qu'est ILP. Le regroupement en locuteurs à base de graphes est un concept récent, ayant été abordé par différents auteurs tels que [Bredin et al., 2014; Campbell et Singer, 2012; Dupuy et al., 2014b; Karam et Campbell, 2013; Shum et al., 2013; van Leeuwen et Brümmer, 2014] (la plupart de ces travaux se sont déroulés en parallèle des nôtres).

Les N classes candidates à un regroupement, ainsi que les distances entre ces classes, peuvent être représentées sous la forme d'un graphe non orienté complet d'ordre N, donc connexe, dans lequel les sommets représentent les classes, et les arêtes les distances. Une représentation schématique illustrant ce concept avec un ensemble de 13 classes candidates est présentée en figure 5.13. L'approche propo-

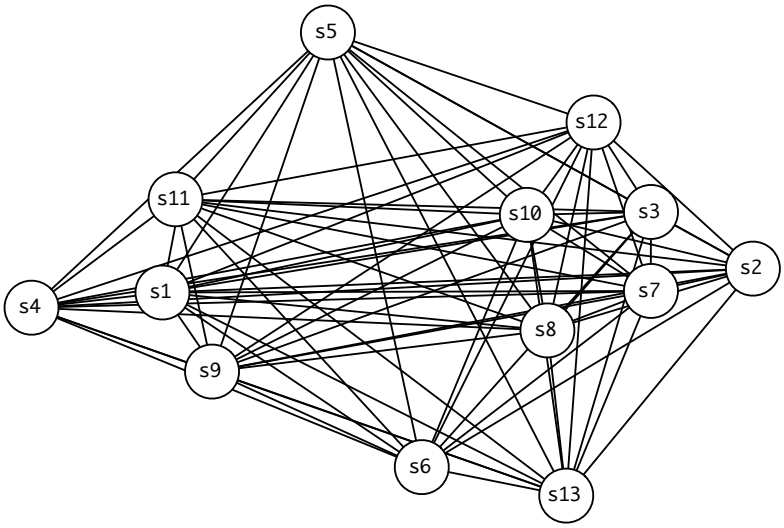

Figure *5.13 – Graphe non orienté complet d'ordre 13, dans lequel les sommets représentent les classes, et les arêtes les distances.*

sée repose sur deux recherches successives : tout d'abord, simplifier le graphe en retirant les arêtes superflues et déterminer les composantes connexes de ce graphe, s'il y en a. Ensuite, rechercher parmi les composantes connexes celles présentant les caractéristiques d'une étoile (graphe biparti complet K_1, n).

▷ Recherche des composantes connexes

Dans un regroupement ILP, si le score entre deux classes c_i et c_j est supérieure à une valeur seuil β [3], alors c_i et c_j ne pourront pas être regroupées ensemble. Par conséquent, les arêtes du graphe connexe (*cf.* figure 5.13) pour lesquelles $d(c_i, c_j) > \beta$ sont superflues et peuvent être retirées sans conséquence. La suppression de ces arêtes permet de mettre en évidence des composantes connexes, illustrées dans le schéma de la figure 5.14, qui correspondent à des sous-problèmes de regroupement indépendants les uns des autres (car désormais, aucune arête ne les relie).

Ces composantes connexes peuvent être déterminées par l'algorithme de parcours en profondeur (Depth First Search – DFS). Cet algorithme, dont le temps d'exécution est polynomial, permet à partir d'un sommet donné d'en déterminer les sommets connexes en $\mathcal{O}(n^2)$. Les composantes connexes sont finalement établies après que chacun des sommets ait été parcouru. La décomposition en composantes connexes, qui permet de constituer des sous-problèmes de regroupement indépendants, aide

3. Nous faisons varier deux seuils dans la partie expérimentale : le seuil β, pour la décomposition en composantes connexes, et le seuil δ, pour les méthodes de classification ILP et HAC.

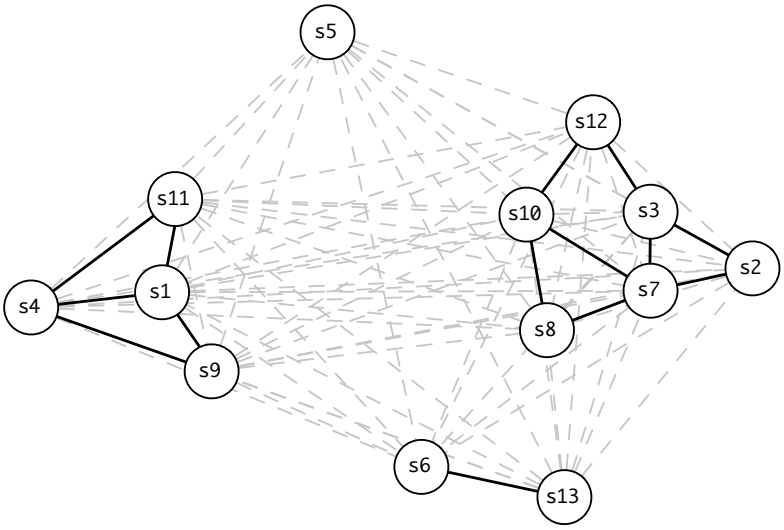

Figure *5.14 – 4 composantes connexes déterminées par l'algorithme de parcours en profondeur après retrait des arêtes superflues.*

à diminuer la complexité du regroupement : il est plus rapide, d'un point de vue algorithmique, de résoudre plusieurs sous-problèmes de regroupement de petite taille plutôt qu'un unique problème de grande taille, même avec l'approche HAC qui est en $\mathcal{O}(n^3)$. De plus, les sous-problèmes peuvent être résolus en parallèle du fait de leur indépendance.

La complexité du problème de regroupement, déjà diminuée par la décomposition du graphe initial en composantes connexes, peut être encore réduite par la recherche des cas triviaux. Ces cas triviaux correspondent aux composantes connexes pour lesquelles la classe centrale est évidente, et cette recherche peut être menée en déterminant quelles composantes présentent les caractéristiques d'un graphe biparti complet K_1, n (ou *étoile*).

▷ Recherche des composantes connexes de type « étoiles »

Les composantes connexes correspondant à des problèmes de regroupement triviaux sont caractérisées par leur aspect en étoile. Une étoile est un graphe biparti complet K_1, n, c'est-à-dire un graphe connexe dont n sommets de profondeur 1 sont connectés à un sommet central K (*cf.* figure 5.15).

L'algorithme permettant de déterminer si une composante connexe présente les caractéristiques d'une étoile est réalisé en $\mathcal{O}(n^2)$, où n est le nombre de classes de la composante connexe pour lesquelles le sommet correspondant est relié à $n-1$ sommets. Si la structure d'une composante connexe présente les caractéristiques

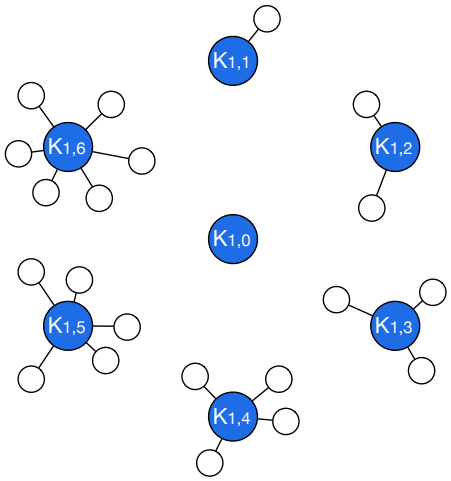

Figure *5.15 – Exemple de graphes biparti complets* K_1, n *avec* $n \in [0, 6]$.

d'une étoile, alors le sous-problème de regroupement correspondant à cette composante connexe est résolu : les classes correspondant aux n sommets de profondeur 1 doivent être regroupées au sein de la classe correspondant au sommet central K. Il est important de souligner le cas particulier de l'étoile $K_1, 0$. Ce cas de figure, présenté dans les figures 5.15 et 5.16, correspond à un sommet isolé pour lequel la classe correspondante n'est pas sujette à regroupement (cas d'un locuteur qui n'est présent que dans un seul enregistrement de la collection). L'étiquette de la classe correspondant au sommet central sera, par la suite, utilisée pour identifier les éléments de la composante connexe.

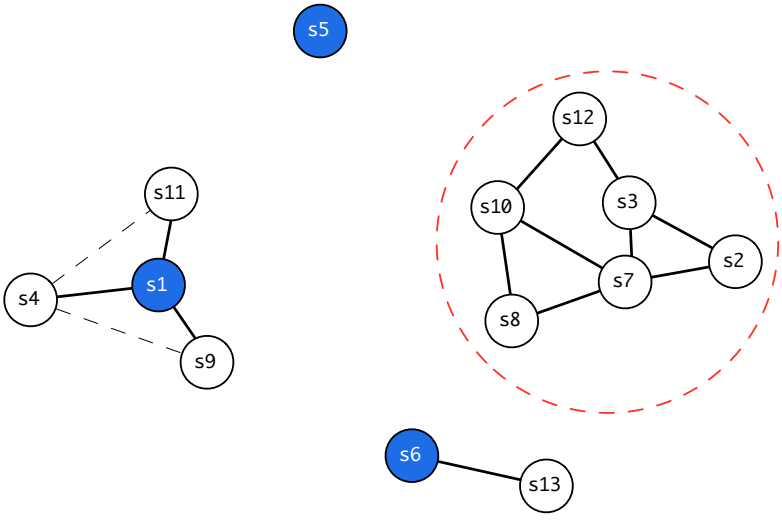

Figure *5.16 – Composantes connexes en* étoile *(sommet central coloré en bleu) et composante connexe complexe (entourée en rouge).*

Lorsque la composante connexe ne présente pas les caractéristiques d'une étoile, elle est dite « complexe ». Les sous-problèmes de regroupement correspondant aux

composantes connexes complexes doivent, idéalement, être résolus par l'approche de regroupement ILP. Rien n'empêche cependant de recourir à d'autres algorithmes de classifications pour traiter les composantes connexes complexes, étant donné l'indépendance des sous-problèmes de regroupement.

Dans l'exemple présenté en figure 5.16, seule la composante entourée en traits interrompus rouges requiert l'utilisation d'un algorithme de regroupement pour résoudre le sous-problème associé. Cette approche permet de simplifier la complexité de la recherche de notre problème de regroupement, de manière exhaustive, uniquement par des algorithmes polynomiaux. Dans le meilleur des cas, toutes les composantes connexes sont des étoiles et le problème de regroupement est résolu sans le moindre recours aux algorithmes de regroupement en locuteurs.

▷ Évaluation et discussion

En termes de composantes connexes

Afin d'illustrer l'intérêt de notre approche de simplification par la théorie des graphes, nous présentons dans le tableau 5.9 une analyse portant sur le nombre et la nature des composantes connexes présentes dans les 7 collections du niveau *programme*. Cette analyse a été menée sur la base d'un regroupement global de type ILP_{PLDA} avec un seuil PLDA $\delta = -30$ (Il s'agit du seuil δ permettant d'obtenir les meilleurs $\text{DER}_{\text{de collections}}$ par regroupement ILP_{PLDA} sur l'ensemble des collections, *Planète Showbiz* exclue). Pour cette expérience, le seuil β de décomposition en composantes connexes (CC) a également été fixé à -30. Le regroupement ILP et l'approche de classification par décomposition en composantes connexes optimisent les mêmes critères, ainsi, lorsque $\delta = \beta$, les résultats en termes de DER sont identiques entre l'approche de classification ILP et l'approche de décomposition en composantes connexes suivie d'une classification ILP (CC+ILP). Dans le tableau 5.9 sont présentés, pour chacune des collections étudiées :

- Le nombre d'enregistrements qui compose les collections (n^{bre} Enr.).
- Le nombre de classes dans la segmentation initiale, qui correspond à la concaténation des segmentations obtenues au niveau *émission* de l'architecture (n^{bre} c. initiales).
- Le nombre total de composantes connexes pour le seuil $\beta = \delta = -30$ (n^{bre} CC. total).
- Le nombre de composantes connexes de type *étoiles* parmi la totalité des composantes connexes (n^{bre} CC. en étoile), avec une distinction entre les étoiles

particulières correspondant à des sommets isolés ($K_1, 0$), et les étoiles plus élaborées pour lesquelles au moins un sommet est attaché au sommet central (K_1, n).

- Le nombre de composantes connexes *complexes*, qui font l'objet d'un regroupement ILP ou HAC (n^{bre} CC. complexes).

Collection	n^{bre} Enr.	n^{bre} c. initiales	n^{bre} CC. total	n^{bre} CC. en étoile $K_1, 0$	n^{bre} CC. en étoile K_1, n	n^{bre} CC. complexes
BFM Story	48	2845	2130	1811 (85,02%) [63,66%]	306 (14,37%) [30,65%]	13 (0,61%) [5,69%]
Planète Showbiz	160	4224	2885	2572 (89,15%) [60,89%]	282 (9,77%) [25,45%]	31 (1,08%) [13,66%]
Ça vous Regarde	23	814	681	626 (91,92%) [76,90%]	54 (7,93%) [20,15%]	1 (0,15%) [2,95%]
Entre les Lignes	27	732	454	395 (87%) [53,96%]	53 (11,67%) [32,65%]	6 (1,33%) [13,39%]
LCP Info	48	1812	1252	1111 (88,74%) [61,31%]	126 (10,06%) [24,59%]	15 (1,20%) [14,13%]
Pile et Face	33	868	612	559 (91,34%) [64,40%]	48 (7,84%) [21.43%]	5 (0,82%) [14,17%]
Top Questions	35	1000	603	524 (86,90%) [52,40%]	67 (11,11%) [26%]	12 (1,99%) [21,60%]
Moyenne	-	**1756,43**	**1231**	**1085,43 (88,18%) [61,80%]**	**133,71 (10,86%) [26,36%]**	**11,86 (0,96%) [11,84%]**

Table *5.9 – Nombre et nature des composantes connexes obtenues sur les sept collections du niveau* programme *par regroupement global* ILP_{PLDA} *avec* $\beta = \delta = -30$. *Les valeurs entre parenthèses représentent les proportions de composantes connexes par rapport au nombre total de composantes connexes. Les valeurs entre crochets représentent les proportions de classes par rapport au nombre total de classes.*

Dans le tableau 5.9, nous présentons également pour chaque collection et chaque type de composante connexe :

- La proportion que représente le nombre de composantes connexes en étoile et complexes par rapport au nombre total de composantes connexes (valeurs entre parenthèses).

- La proportion de classes de locuteurs impliquées dans les composantes connexes de type *étoiles* et complexes par rapport au nombre total de classes de locuteurs (valeurs entre crochets).

En moyenne, sur les sept collections étudiées, 88,18% des composantes connexes correspondent à des sommets isolés. Ces sommets représentent les classes de locuteurs pour lesquelles aucun regroupement n'est possible étant donné le seuil β. Aucun algorithme de regroupement n'est donc nécessaire. 10,86% des composantes connexes présentent les caractéristiques d'une étoile impliquant au moins deux sommets. Les classes correspondant aux n sommets attachés au sommet central K sont alors regroupées au sein de la classe représentée par le sommet central K. Les composantes connexes de type *étoiles* (cas particulier des sommets isolés inclus) représentent donc 99,04% de la totalité des composantes connexes. Les composantes connexes complexes, dont la solution de regroupement n'est pas évidente, ne représentent que 0,96% de la totalité des composantes connexes. Cela représente, en moyenne par collection, 12 sous-problèmes indépendants devant être traités par une approche de regroupement, en l'occurrence dans le cas présent, un regroupement ILP dans sa configuration PLDA avec un seuil δ fixé à -30. En moyenne toujours, 61,80% des classes initiales sont représentées par des sommets isolés. Les étoiles (au moins deux sommets) représentent 26,36% du nombre de classes initiales, et les composantes complexes connexes, seulement 11,84%. En conclusion, la décomposition en composantes connexes permet de simplifier un unique problème de regroupement constitué de 1756 classes, en moyenne, à seulement 12 sous-problèmes indépendants constitués d'environ 18 classes chacun.

En termes de $\text{DER}_{\text{de collections}}$

Nous présentons dans le tableau 5.10 les résultats en termes de $\text{DER}_{\text{de collections}}$ obtenus sur les collections du niveau *programme* avec les approches de regroupement ILP_{PLDA} et HAC_{PLDA}, avec et sans recours à la simplification par décomposition en composantes connexes. La collection *Planète Showbiz* n'est pas étudiée, car trop atypique (*cf.* résultats présentés en partie 5.2.2). Afin de jauger l'importance du couplage entre la décomposition en composantes connexes (*CC.* dans le tableau) et les approches de regroupement HAC et ILP, nous avons fixé le seuil β à -30. Nous présentons également dans le tableau 5.10 les taux d'erreur $\text{DER}_{\text{de collections}}$ obtenus par la seule décomposition en composantes connexes. La stratégie mise en œuvre pour gérer le cas des composantes connexes complexes sans recourir aux approches de regroupement citées est simple : les classes impliquées dans une composante connexe complexe sont toutes regroupées au sein d'une unique classe (colonne *CC. + Fusion* dans le tableau 5.10).

Configuration	ILP	CC. + ILP	HAC	CC. + HAC	CC. + Fusion
Seuil β	-	-30	-	-30	-30
Seuil δ	-30	-30	30	30	-
BFM Story	15,03%	15,03%	14,12%	14,72%	16,50%
Ça vous Regarde	21,82%	21,82%	21,65%	21,82%	21,82%
Entre les Lignes	15,34%	15,34%	15,16%	15,01%	23,32%
LCP Info	21,30%	21,30%	18,65%	20,20%	22,03%
Pile et Face	23,74%	23,74%	23,94%	23,74%	23,36%
Top Questions	29,24%	29,24%	33,57%	31,28%	47,48%
Moyenne	**19,43%**	**19,43%**	**19,08%**	**19,35%**	**23,30%**

Table *5.10 – Comparaison des résultats en termes de* $\mathbf{DER_{de\ collections}}$ *sur les collections du niveau* Programme, Planète Showbiz *exclue, pour les regroupements HAC et ILP, avec et sans application de la décomposition en composantes connexes pour un seuil* β *égal à -30.*

Les moyennes présentées ont été calculées en fonction de la durée évaluée pour des collections étudiées, sans tenir compte de la collection *Planète Showbiz*. Les résultats présentés dans les colonnes *ILP* et *HAC* sont ceux déjà présentés dans la partie précédente. Leurs seuils δ (respectivement -30 et 30) sont ceux ayant permis d'atteindre les taux d'erreur les plus faibles sans recourir à l'approche de décomposition en composantes connexes. Ces résultats représentent en quelque sorte les valeurs étalons de nos expériences. Les résultats obtenus avec les approches de regroupement *ILP* (δ = -30) et *CC+ILP* ($\delta = \beta$ = -30) sont en tout point identiques (19,43% en moyenne), ce qui n'est pas surprenant en soi étant donné que l'approche de décomposition en composantes connexes a été pensée de manière à simplifier la complexité du regroupement ILP. Le constat est différent pour l'approche de regroupement HAC : la décomposition en composantes connexes dégrade légèrement les résultats comparé à la seule utilisation du regroupement HAC (-0,27% en moyenne), cependant, le $\text{DER}_{\text{de collections}}$ moyen ainsi obtenu reste inférieur à celui des approches de regroupement *ILP* et *CC+ILP* (19,35% *vs.* 19,43%). On notera cependant que ce gain est très faible. Enfin, la seule décomposition en composantes connexes (colonne *CC+Fusion*) permet d'atteindre un $\text{DER}_{\text{de collections}}$ égal à 23,30% en moyenne. En observant le détail sur les différentes collections, on constate cependant que l'approche *CC+Fusion* permet d'atteindre des $\text{DER}_{\text{de collections}}$ inférieurs ou égaux à ceux des autres approches étudiées pour la collection *Pile et Face* (23,36%).

La valeur des seuils β, pour la décomposition en composantes connexes, et δ, pour l'approche de regroupement *CC+ILP*, devrait toujours être identique, car le critère optimisé est le même : le rayon des classes. S'il est concevable que la valeur de β puisse être supérieure à celle de δ, le contraire n'aurait que peu de sens : le regroupement ILP serait alors effectué sur des composantes connexes complexes dont

la distance entre les classes serait inférieure à la valeur de δ. Concernant l'approche de regroupement *CC+HAC*, employer des valeurs différentes pour les seuils β et δ afin d'optimiser les résultats en termes de $\text{DER}_{\text{de collections}}$ a du sens, car le critère d'optimisation n'est pas le même : l'approche de regroupement HAC optimise le diamètre des classes, du fait du critère de liaison, qui correspond au saut maximum.

Nous présentons dans le tableau 5.11 les résultats obtenus avec les combinaisons de seuils β et δ permettant d'obtenir les meilleurs taux d'erreur $\text{DER}_{\text{de collections}}$, en moyenne, sur les collections du niveau *programme* (sans tenir compte de la collection *Planète Showbiz*).

Configuration	**ILP**	**CC. + ILP**	**CC. + ILP**	**HAC**	**CC. + HAC**	**CC. + Fusion**
Seuil β	-	-30	20	-	20	-50
Seuil δ	-30	-30	-40	30	0	-
BFM Story	15,03%	15,03%	13,72%	14,12%	13,92%	15,26%
Ça vous Regarde	21,82%	21,82%	21,75%	21,65%	21,97%	21,82%
Entre les Lignes	15,34%	15,34%	15,16%	15,16%	15,16%	14,50%
LCP Info	21,30%	21,30%	18,37%	18,65%	18,53%	23,24%
Pile et Face	23,74%	23,74%	23,94%	23,94%	23,94%	23,74%
Top Questions	29,24%	29,24%	28,53%	33,57%	29,76%	30,69%
Moyenne	**19,43%**	**19,43%**	**18,33%**	**19,08%**	**18,60%**	**20,07%**

Table *5.11 – Comparaison des résultats en termes de* $\textbf{DER}_{\textbf{de collections}}$ *sur les collections du niveau* Programme, Planète Showbiz *exclue, pour les regroupements HAC et ILP, avec et sans application de la décomposition en composantes connexes. Les résultats présentés correspondent aux meilleurs résultats atteignables en moyenne (collection* Planète Showbiz *exclue).*

Le couplage des approches de décomposition en composantes connexes avec les approches de regroupement ILP et HAC nous a permis d'atteindre des taux d'erreur $\text{DER}_{\text{de collections}}$ inférieurs à ceux obtenus sans l'approche de décomposition, en jouant sur la valeur du seuil β et en adaptant la valeur du seuil δ en conséquence. Le seuil β influence directement le nombre et la nature des composantes connexes, il correspond à la valeur maximale du score de vraisemblance pour laquelle deux classes peuvent être regroupées au sein d'une même composante connexe. Un seuil β fixé à 30 va permettre de concevoir des composantes connexes où le score PLDA entre deux classes est inférieur ou égal à 30. Le nombre de composantes connexes de type *étoile* (sommets isolés et étoiles élaborées), et le nombre de composantes connexes complexes, varient donc en fonction de β. Concernant les composantes connexes complexes, qui nécessitent l'application d'une approche de regroupement de type HAC ou ILP pour être idéalement classifiées, le nombre de sommets impliqués évolue également en fonction de β. Plus la valeur de β est élevée, plus le nombre de classes impliquées dans les composantes connexes complexes est élevé (en revanche,

le nombre de composantes connexes complexes diminue). La valeur de δ est donc susceptible d'évoluer en fonction du seuil β.

Les résultats présentés dans les colonnes *CC+ILP* et *CC+HAC* correspondent aux meilleurs taux d'erreur $\text{DER}_{\text{de collections}}$ que nous avons pu obtenir en faisant varier empiriquement β et δ. Ces $\text{DER}_{\text{de collections}}$ sont meilleurs que ceux obtenus par les mêmes approches de regroupement sans recourir à la décomposition en composantes connexes : l'approche *CC+HAC* permet de gagner en moyenne 0,48%, en absolu, par rapport à l'approche *HAC*, et l'approche *CC+ILP* permet quant à elle de réduire le $\text{DER}_{\text{de collections}}$, en moyenne, de 1,10% en absolu, par rapport à l'approche *ILP* (18,33% contre 19,43%). Comparé à ces résultats, l'approche *simpliste* mise en œuvre pour gérer les composantes connexes complexes en s'affranchissant des approches de regroupement n'est pas aussi efficace. L'approche *CC+Fusion* permet néanmoins d'approcher les $\text{DER}_{\text{de collections}}$ obtenus par les approches *HAC* et *ILP*, avec un DER moyen égal à 20,07%. En revanche, les résultats obtenus sur certaines collections, en particulier *Pile et Face* et *Entre les Lignes*, sont meilleurs ou égaux à ceux des approches de implémentant les regroupements ILP et HAC. L'approche *CC+Fusion*, uniquement basée sur la décomposition en composantes connexes, pourrait se révéler intéressante dans un cadre applicatif où la rapidité de traitement primerait sur la qualité de la segmentation produite, ou si les ressources disponibles ne seraient pas suffisamment conséquentes, comme sur un terminal mobile (smartphone, tablette, *etc.*).

Notre approche de simplification par décomposition en composantes connexes est très proche de l'approche proposée par [Campbell et Singer, 2012], qui repose sur la méthode des *K plus proches voisins*. Elle est aussi rapide à calculer sauf que nous avons recours au saut maximum pour traiter les composantes connexes complexes, avec la méthode de regroupement HAC, contrairement à [Campbell et Singer, 2012], qui a recours à une méthode de saut minimum, introduisant ainsi des effets de chaînage.

Discussion

Nous avons obtenu de meilleurs résultats avec les approches *CC+ILP* et *CC+HAC* lorsque la valeur du seuil β est supérieure à celle du seuil δ. Plus le seuil β est élevé, plus la classification résultant de la décomposition en composantes connexes est grossière, et plus les approches de regroupement HAC ou ILP sont sollicitées pour affiner cette classification. En effet, en fixant un seuil β élevé nous obtenons une première classification où les composantes connexes de type étoile les plus évidentes sont identifiées. De ce fait, les composantes connexes complexes sont plus denses,

mais les décisions de classification les plus « difficiles » ne seront prises que par les algorithmes de classification. En fixant un seuil δ inférieur à β, ce sont donc les approches de regroupement ILP et HAC qui prennent les décisions de classification difficiles, permettant ainsi d'affiner la classification des composantes connexes complexes.

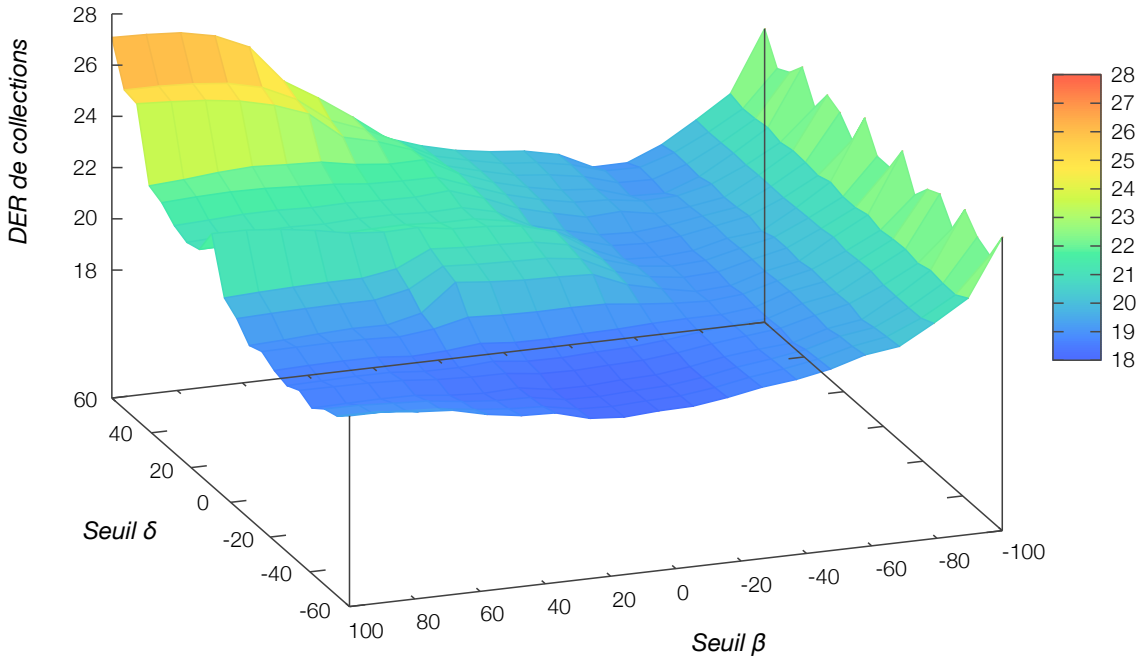

Figure *5.17 – Surface représentant le taux $DER_{\text{de collections}}$ moyen obtenu en fonction des seuils β et δ sur les 6 collections de niveau* programme *étudiées pour l'approche* CC+ILP.

S'il n'est pas surprenant de constater des améliorations lorsque β est supérieur à δ, il est difficile de déterminer une combinaison de seuils *stable*, en moyenne, avec nos collections dont le type des enregistrements (débats, journaux, *etc.*) diffère. Nous présentons à cet effet, en figure 5.17, une représentation graphique de l'évolution du $\text{DER}_{\text{de collections}}$ en fonction de différentes valeurs pour les seuils β et δ pour la méthode de classification *CC+ILP*.

Il n'est pas naturel de fixer un seuil β pour la décomposition en composantes connexes inférieur à celui du seuil δ de l'approche de regroupement, en particulier pour l'approche de regroupement ILP, puisque le critère optimisé est le même : le rayon des classes (contrairement à l'approche de regroupement HAC qui optimise le diamètre). On peut pourtant constater, à l'aide de la figure 5.17, qu'il est possible d'atteindre des $\text{DER}_{\text{de collections}}$ très proches de la valeur optimale lorsque β est inférieur à δ, en particulier pour les plus petites valeurs de β testées. Ce phénomène est illustré dans le tableau 5.12, où nous présentons les $\text{DER}_{\text{de collections}}$ compris entre les valeurs *normale* (19,43%, avec $\beta = \delta$) et *optimale* (18,33%, avec $\beta > \delta$) (*cf.* tableau 5.11) lorsque β est inférieur à δ pour l'approche CC+ILP, avec β = -60.

Approche		Seuil β	Seuil δ	$DER_{\text{de collections}}$
CC+ILP	$\beta = \delta$	-30	-30	19,43%
	$\beta > \delta$	20	-40	18,33%
	$\beta < \delta$	-60	-20	19,34%
		-60	-10	18,89%
		-60	0	18,92%
		-60	10	18,94%
		-60	20	18,69%
		-60	30	19,08%

Table 5.12 – *Exemple de* $\mathbf{DER_{\text{de collections}}}$ *obtenus par l'approche de regroupement* CC+ILP *lorsque* β *est inférieur à* δ*, avec* β *= -60.*

Une observation similaire peut être établie avec le regroupement *CC+HAC*, dont la représentation graphique de l'évolution du $DER_{\text{de collections}}$ en fonction de différentes valeurs pour les seuils β et δ est présentée en figure 5.18. Ce phénomène est cependant moins présent et moins dérangeant, dans la mesure où le critère optimisé par l'approche HAC (diamètre des classes) est différent de celui de la décomposition en composantes connexes (rayon des classes).

Nous expliquons ce phénomène par le fait qu'en moyenne seulement 1% des composantes connexes sont complexes, et donc la classification réalisée par ILP ou HAC n'a que peu d'impact sur la classification finale : nous avons constaté avec l'approche *CC+Fusion*, où les classes d'une composante connexe complexe sont regroupées au sein d'une classe unique, que les approches de regroupement HAC et ILP ne sont pas nécessairement utiles à la classification : le $DER_{\text{de collections}}$ obtenu avec l'approche CC+Fusion sur la collection *Pile et Face* (23,36%), est inférieur de 0,38% aux $DER_{\text{de collections}}$ obtenu avec les approches CC+ILP et CC+HAC (*cf.* tableau 5.11).

En outre, les résultats présentés en fonction des seuils β et δ ne correspondent qu'aux moyennes des $DER_{\text{de collections}}$ observés sur six collections différentes. Pour une collection donnée, les approches de regroupement CC+ILP et CC+HAC donnent des résultats très proches, la différence n'est que de 1% (plus ou moins). Le choix de la méthode de classification ne devrait donc pas reposer sur le DER, trop peu dissemblable entre HAC et ILP, mais sur un choix algorithmique ou technique :

- *CC+HAC* : algorithme de classification en temps polynomial, mais deux critères différents sont à optimiser.
- *CC+ILP* : algorithme non polynomial, mais plus élégant, car un seul et même critère est à optimiser.

Enfin, il faut également prendre en compte le fait que seule une portion des don-

nées d'expérimentation sont annotées à des fins d'évaluation (67 heures sur les 178 heures d'enregistrements audio), et les moyennes des $DER_{\text{de collections}}$ sont obtenues en pondérant, pour chaque collection, les résultats par la durée évaluée.

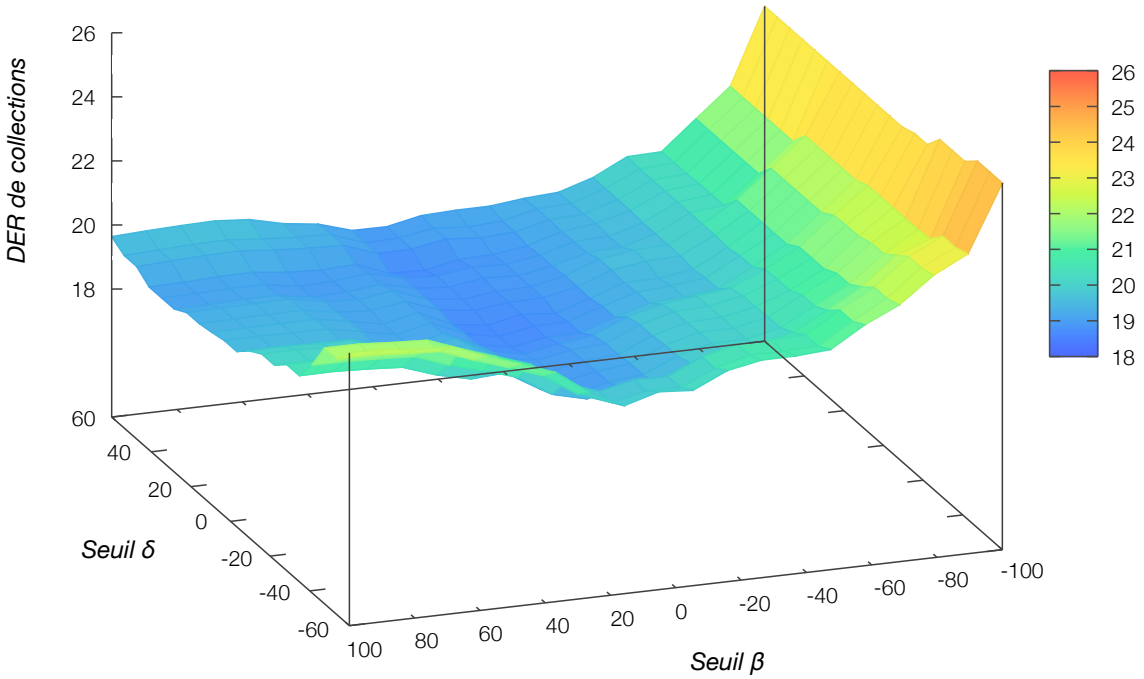

Figure *5.18 – Surface représentant le taux $DER_{\text{de collections}}$ moyen obtenu en fonction des seuils β et δ sur les 6 collections de niveau* programme *étudiées pour l'approche* CC+HAC.

Nous présenterons une série de résultats expérimentaux à l'issue de ce chapitre. Afin de conserver une certaine cohérence, nous avons fait le choix de présenter les résultats obtenus par l'approche *CC+ILP* pour une combinaison de seuils $\beta = \delta$. Nous présenterons, de ce fait, les résultats obtenus par l'approche *CC+HAC* avec la même valeur pour le seuil β.

Bilan

L'approche de simplification présentée dans cette partie, qui repose sur la décomposition en composantes connexes d'un graphe complet, permet de rechercher les séparations évidentes en temps polynomial. Cette approche de simplification, qui permet finalement de minimiser le travail des approches de classification ILP et HAC, apporte un gain, mais rend les justifications théoriques plus bancales.

5.3. Regroupements intra-émission

Une question s'est posée quant à la manière d'effectuer le regroupement global du niveau *collection* de l'architecture. En effet, deux stratégies antagoniques, déjà évoquées lors de la présentation de notre architecture dans la partie 5.1, peuvent être considérées quant au comportement des approches de regroupement vis-à-vis des classes issues d'un même enregistrement. Étant donné que le regroupement global du niveau *collection* porte sur la réunion des segmentations obtenues au niveau *émission* de l'architecture par le système de SRL d'émissions, deux hypothèses peuvent être envisagées. Pour rappel :

1. Les segmentations produites au niveau *émission* sont imparfaites et peuvent être améliorées grâce au regroupement global du niveau *collection*. Dans ce cas, la liberté de regrouper des classes issues d'un même enregistrement doit être laissée aux approches de regroupement employées au niveau *collection*.
2. Les segmentations produites au niveau *émission* sont considérées comme *idéales*, auquel cas le regroupement global du niveau *collection* ne doit pas les altérer. Il convient donc d'empêcher les approches de regroupement global de regrouper des classes issues d'un même enregistrement.

5.3.1 Autoriser les regroupements intra-émission

Étant donné les taux d'erreur $\text{DER}_{\text{d'émissions}}$ obtenus sur chacun des enregistrements qui composent une collection (rarement nuls), il est tout à fait concevable de considérer que les segmentations produites au niveau *émission* sont imparfaites et pourraient, grâce au regroupement du niveau *collection*, être améliorées. L'hypothèse sous-jacente est que les locuteurs d'un enregistrement ne sont pas forcément représentés par une unique classe dans la segmentation de niveau *émission* correspondante. Deux ou plusieurs classes représentant un même locuteur pourraient ainsi être regroupées grâce à une classe « pivot » provenant d'un enregistrement différent. Le regroupement global du niveau *collection* permettrait ainsi d'améliorer les regroupements effectués dans les segmentations du niveau *émissions*. Dans ce cas, il convient de laisser la liberté aux approches de regroupement du niveau *collection* de regrouper les classes, quel que soit l'enregistrement dont elles sont issues. Cette stratégie ne permet cependant pas de corriger les erreurs correspondant à des regroupements excessifs (cas où la classe d'un locuteur serait injustement regroupée avec celle d'un locuteur différent), ni celles dues à une mauvaise segmentation. Un exemple illustrant l'intérêt de cette stratégie est présenté en figure 5.19 :

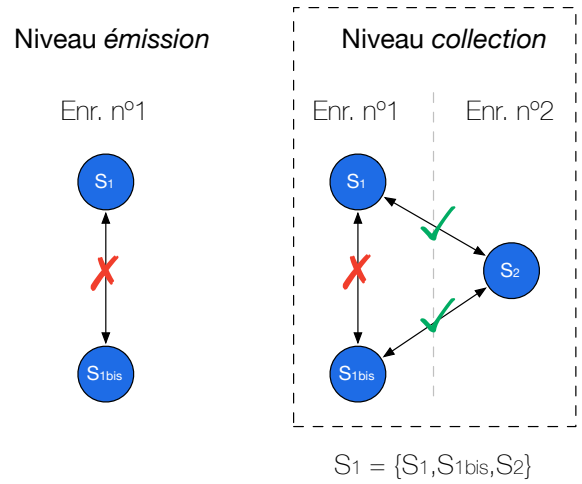

Figure *5.19 – Stratégie de regroupement autorisant le regroupement de deux classes issues d'un même enregistrement.*

Dans cet exemple, les deux classes S_1 et S_{1bis} de l'enregistrement n°1 sont censées représenter un même locuteur. Or, il est établi au niveau *émission* que la similarité entre ces deux classes est supérieure à la valeur seuil δ. Ces deux classes ne sont donc pas regroupées, générant ainsi une erreur de confusion lors du calcul du taux DER. En revanche, au niveau *collection*, si les similarités entre ces deux classes et une classe tierce S_2 représentant le même locuteur dans un second enregistrement sont toutes les deux en deçà du seuil δ, alors un regroupement global peut être effectué entre ces trois classes. L'erreur de regroupement du niveau *émission* impliquant les deux classes S_1 et S_{1bis} est alors corrigée au niveau *collection* par l'intermédiaire de la classe « pivot » S_2.

Cette stratégie correspond au comportement naturel des approches de regroupement HAC et ILP, telles qu'héritées de la SRL d'émissions. En effet, le regroupement global du niveau *collection* se comporte de la même manière que le regroupement du niveau *émission* : chacune des classes impliquées dans le regroupement peut être regroupée, quel que soit leur enregistrement de provenance, jusqu'à former une unique classe si la valeur du seuil δ est élevée. Implémenter cette stratégie ne nécessite donc aucune modification dans les approches de regroupement, contrairement à la stratégie alternative présentée dans la section suivante.

5.3.2 Empêcher les regroupements intra-émission

La deuxième façon de voir les choses est de considérer que les segmentations produites au niveau *émission* sont les plus abouties possible et que, par conséquent, le regroupement global du niveau *collection* ne doit pas les altérer. Selon cette hy-

pothèse, les locuteurs d'un enregistrement sont correctement représentés par une unique classe dans la segmentation de niveau *émission* correspondante. Le regroupement de niveau *collection* ne porterait donc que sur des classes provenant exclusivement d'émissions différentes. Il convient donc, dans ce cas, de spécifier aux approches de regroupement employées au niveau *collection* qu'il est interdit de regrouper des classes issues d'un même enregistrement. Or, préserver les segmentations établies au niveau *émission* n'est pas une tâche aussi triviale qu'il n'y paraît. Les approches de regroupement actuellement employées sont héritées de la SRL d'émissions, où l'on ne se préoccupe pas de savoir à quelle émission appartient une classe de locuteur puisque les enregistrements sont traités indépendamment les uns des autres. Interdire, au niveau *collection*, le regroupement entre les classes issues d'un même enregistrement peut être réalisé par l'ajout de contraintes (ou règles) de regroupement. Un exemple simple permettant d'illustrer le problème est présenté en figure 5.20 :

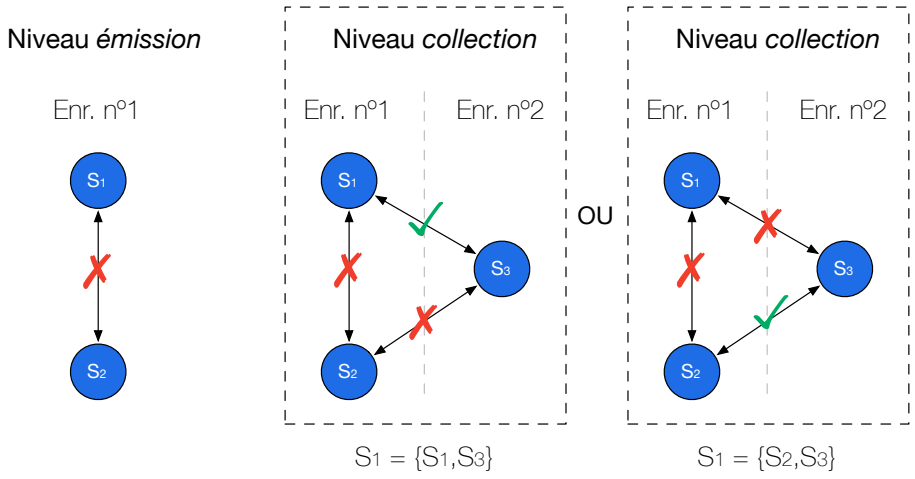

Figure *5.20 – Stratégie de regroupement empêchant le regroupement global de deux classes issues d'un même enregistrement.*

Dans cet exemple, les deux classes S_1 et S_2 de l'enregistrement n°1 sont censées représenter deux locuteurs différents. Le regroupement entre ces deux classes doit donc être interdit au niveau *collection*, quelle que soit leur similarité. Il s'agit de la première des deux contraintes de regroupement évoquées, et la plus simple à faire respecter, car il n'est pas nécessaire de modifier l'algorithme de regroupement. Les scores entre les classes sont calculés en amont du procédé de regroupement. La distance entre deux classes i et j provenant d'un même enregistrement peut être artificiellement augmentée ($d(i,j) \rightarrow +\infty$) de manière à ne jamais satisfaire le critère de regroupement défini par le seuil δ.

Les deux classes S_1 et S_2 peuvent être regroupées avec une classe S_3, issue d'un second enregistrement. Or, si les classes S_1 et S_3 sont regroupées, formant ainsi une classe $\{S_1,S_3\}$, il est nécessaire de s'assurer que la classe S_2 ne puisse pas être

regroupée avec $\{S_1,S_3\}$, sans quoi les deux classes provenant de l'enregistrement n°1 auront été regroupées. En d'autres termes, il faut s'assurer qu'aucune classe « pivot » ne permette de regrouper deux classes issues d'un même enregistrement. Il s'agit de la deuxième contrainte à formuler pour empêcher le regroupement des classes intrinsèques aux enregistrements, et la complexité de son implémentation dépend de l'approche de regroupement employée :

Regroupement HAC : notre approche de regroupement hiérarchique met en œuvre le critère de liaison maximum, où la distance entre deux classes i et j correspond à la plus grande des distances séparant les classes i et j. Aucune modification de l'algorithme de regroupement n'est donc nécessaire étant donné ce critère de liaison, car les distances entre deux classes issues d'un même enregistrement tendent vers $+\infty$ (sous réserve d'avoir préalablement modifié les distances pour satisfaire la première contrainte de regroupement). En revanche, si le critère de liaison utilisé correspond à la distance minimum, ou moyenne, il est nécessaire de modifier l'algorithme de regroupement de manière similaire à ce qui est proposé pour l'approche ILP.

Regroupement ILP : il est nécessaire d'ajouter une contrainte à la définition du problème ILP, en plus de celles déjà présentes. Soit deux classes i et j provenant d'un même enregistrement (avec $d(i,j) \rightarrow +\infty$). La formulation du problème ILP (*cf.* équation 5.1a) permet déjà de ne pas de regrouper i et j, étant donné que nous travaillons sur des sous-ensembles du problème constitués des seuls groupes de classes satisfaisant la contrainte de distance $s(i,j) < \delta$. En revanche, si l'on considère également une classe k (classe « pivot », qui correspond à la classe centrale d'un regroupement ILP), issue d'un autre enregistrement, pour laquelle les distances $s(k,i)$ et $s(k,j)$ seraient toutes deux inférieures à la distance δ, alors une contrainte supplémentaire s'impose (*cf.* figure 5.21).

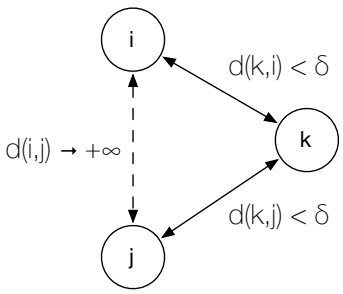

Figure *5.21 – Représentation simplifiée du problème de regroupement ILP pour illustrer le besoin d'une contrainte interdisant les regroupements de classes intrinsèques aux enregistrements.*

En effet, si la variable ILP $x_{k,i}$ est sélectionnée, alors $x_{k,j}$ ne doit pas l'être, et inversement. Nous formulons la contrainte permettant d'empêcher le regroupement des classes i et j par l'intermédiaire d'une classe k de la manière suivante. Soit $C \in \{1 \dots N\}$, le nombre de classes déterminé automatiquement, et $K_{j \in C} = \{k/s(k,j) < \delta\}$ l'ensemble des valeurs possibles de k pour lesquelles les distances entre les classes k et j sont inférieures à la distance δ :

$$x_{k,i} + x_{k,j} - x_{k,k} \leq 0 \qquad k \in K_j, i \in C, j \in C \tag{5.2}$$

où $x_{k,k}$ est la variable binaire égale à 1 si la classe k est sélectionnée pour être centre de classe, et $x_{k,i}$ (respectivement $x_{k,j}$) est égale à 1 quand la classe i (respectivement, j) peut être associée à la classe k. Cette contrainte permet d'empêcher la sélection simultanée des variables $x_{k,i}$ et $x_{k,j}$ lorsque $x_{k,k}$ est désigné pour représenter le centre de la nouvelle classe.

Simplification par la théorie des graphes : le problème est identique à celui soulevé pour le regroupement ILP, une stratégie doit être appliquée afin d'empêcher que deux ou plusieurs classes issues d'un même enregistrement ne soient regroupées en composante connexe par l'intermédiaire d'une classe tierce « pivot ». Les composantes connexes peuvent être réparties en trois catégories selon leurs complexités : les composantes connexes correspondant à des sommets isolés (cas particulier d'une étoile constituée d'un seul sommet), les composantes connexes en étoile impliquant au moins deux sommets, et les composantes connexes complexes (*cf.* partie 5.2.3). Les composantes connexes correspondant aux sommets isolés ne sont pas affectées par la contrainte visant à empêcher le regroupement des classes d'un même enregistrement, puisqu'aucun regroupement n'est possible pour ces classes. Les composantes connexes complexes sont quant à elles gérées par l'une des approches de regroupement présentées précédemment. Reste le cas des composantes connexes en étoile, qui peuvent potentiellement impliquer plusieurs classes d'un même enregistrement. Augmenter artificiellement les distances entre les classes d'un même enregistrement, en amont du procédé de décomposition en composantes connexes, ne résout que partiellement le problème soulevé. Dans l'approche permettant le regroupement des classes d'un même enregistrement, les classes représentées par les sommets d'une étoile sont immédiatement regroupées. Ici, il est nécessaire de vérifier que les sommets constituant une composante connexe en étoile proviennent bien d'enregistrements différents, sans quoi le regroupement n'est pas toléré. Deux cas de figure sont donc à prendre en compte : ou bien les sommets correspondent à des classes issues d'enregistrements différents, auquel cas le regroupement est toléré, ou bien ce n'est pas le cas et la com-

posante connexe étoilée doit être considérée comme une composante connexe complexe, afin d'être traitée par regroupement HAC ou ILP avec les règles permettant aux classes d'un même enregistrement de ne pas être regroupées.

5.3.3 Évaluation et discussion

Laisser la liberté de regrouper des classes issues d'un même enregistrement correspond à la stratégie mise en place pour chacune des expériences présentées jusqu'à présent. Nous proposons dans cette section de comparer les deux stratégies de regroupement global pour les collections. À cet effet, nous présentons les résultats obtenus sur les collections du niveau *programme* en tolérant (stratégie dite « sans contrainte ») et en empêchant (stratégie dite « avec contrainte ») le regroupement des classes intrinsèques aux enregistrements. Afin de faciliter cette comparaison, la collection *Planète Showbiz* n'est pas étudiée (en raison de son caractère atypique par rapport aux autres collections du niveau *programme*), et l'approche de simplification par la théorie des graphes n'est pas employée (en raison de l'influence du seuil β par rapport au seuil de décision δ). La comparaison que nous présentons à été réalisée avec l'approche de regroupement HAC_{PLDA}. Deux raisons motivent ce choix :

1. D'une part, la contrainte visant à empêcher le regroupement des classes issues d'un même enregistrement est très simple à mettre en œuvre et ne modifie pas la complexité de l'algorithme, contrairement à l'approche de regroupement ILP_{PLDA} où la contrainte supplémentaire complexifie le problème et donc, sa résolution. Des tests effectués avec l'approche de regroupement ILP_{PLDA} sur de petites collections montrent des résultats similaires, cependant, un phénomène d'explosion combinatoire survient durant la résolution du problème ILP dès lors que trop de classes sont impliquées, empêchant de ce fait le traitement de collections de taille plus conséquentes en temps raisonnable.
2. D'autre part, l'approche HAC_{PLDA} peut être évaluée sur une plage de seuils δ très large, contrairement à l'approche de regroupement ILP_{PLDA} où les collections les plus volumineuses ne peuvent pas être traitées à partir d'une certaine valeur de δ, à moins de recourir à l'approche de décomposition en composantes connexes (*cf.* partie 5.2.2).

La figure 5.22 présente les résultats, en termes de $\text{DER}_{\text{d'émissions}}$, obtenus pour les deux stratégies de regroupement. Sur cette figure, les résultats obtenus avec la stratégie visant à interdire le regroupement des classes d'un même enregistrement sont représentés, pour chaque collection étudiée, par une ligne en traits interrompus. Conformément au but recherché, les segmentations produites au niveau *émission* de

l'architecture ne sont pas altérées par le regroupement global du niveau *collection*, et les taux d'erreur $\text{DER}_{\text{d'émissions}}$ demeurent identiques, quel que soit le seuil PLDA δ évalué.

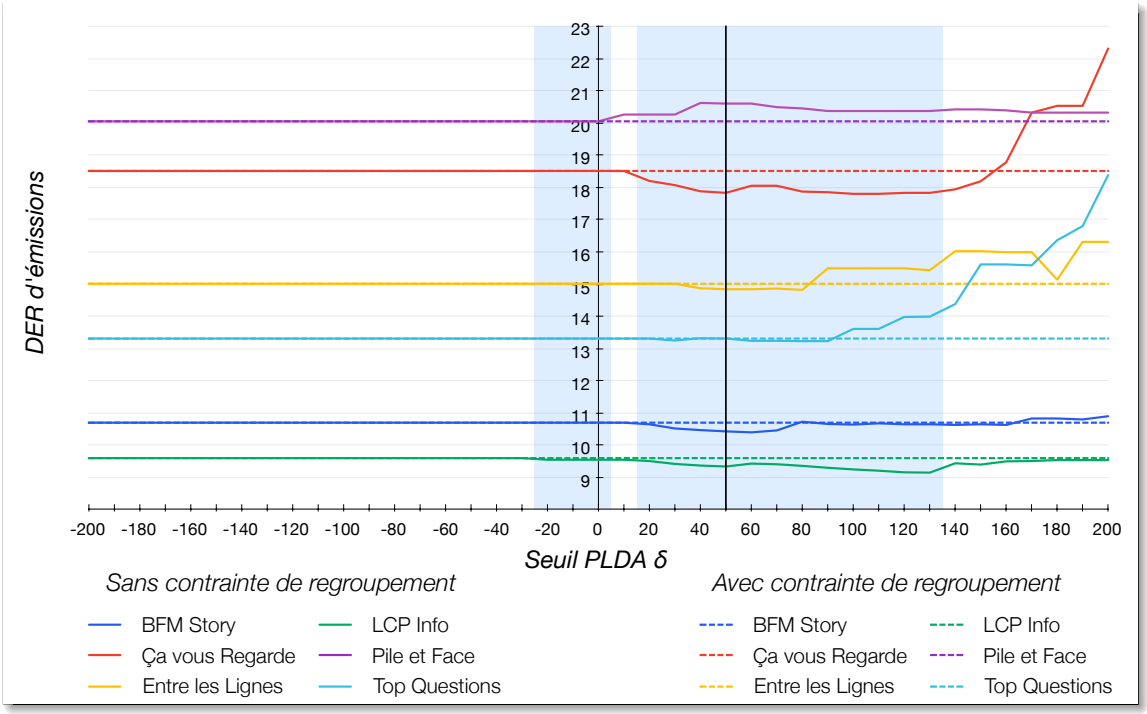

Figure *5.22* – **$\text{DER}_{\text{d'émissions}}$** *obtenus par regroupement global* HAC_{PLDA}*, pour différentes valeurs du seuil PLDA, pour les deux stratégies de regroupement global.*

Les résultats obtenus avec la stratégie opposée, permettant le regroupement des classes d'un même enregistrement, sont représentés par les courbes pleines (déjà présentés dans la partie 5.2.2). On observe une légère amélioration des taux d'erreur $\text{DER}_{\text{d'émissions}}$, pour toutes les collections étudiées excepté *Pile et Face*, pour une certaine valeur de δ (entre 10 et 80, en général). Il n'y a cependant pas de tendance générale, les variations observées dépendent de la collection étudiée et du seuil δ. Toutefois, en moyenne sur ces collections, ces taux $\text{DER}_{\text{d'émissions}}$ sont légèrement inférieurs à ceux obtenus lorsque la stratégie « avec contrainte » est appliquée (aires bleues sur la figure 5.22), avec un minimum atteignant 12,98% pour le seuil $\delta = 50$, contre 13,17% dans le cas contraire.

Les résultats en termes de $\text{DER}_{\text{de collections}}$ sont présentés de manière similaire en figure 5.23, avec, en traits interrompus, les résultats correspondant à la stratégie de regroupement « avec contrainte », et en traits pleins, les résultats correspondant à la stratégie de regroupement « sans contrainte ». Sur cette figure comme sur la précédente, on constate que les résultats sont identiques entre les deux stratégies de regroupement global (les courbes pleines et en traits interrompus se chevauchent) jusqu'à $\delta = -30$. Cette particularité signifie que les regroupements globaux effectués

jusqu'à cette valeur de δ n'impliquent aucune classe provenant d'un même enregistrement.

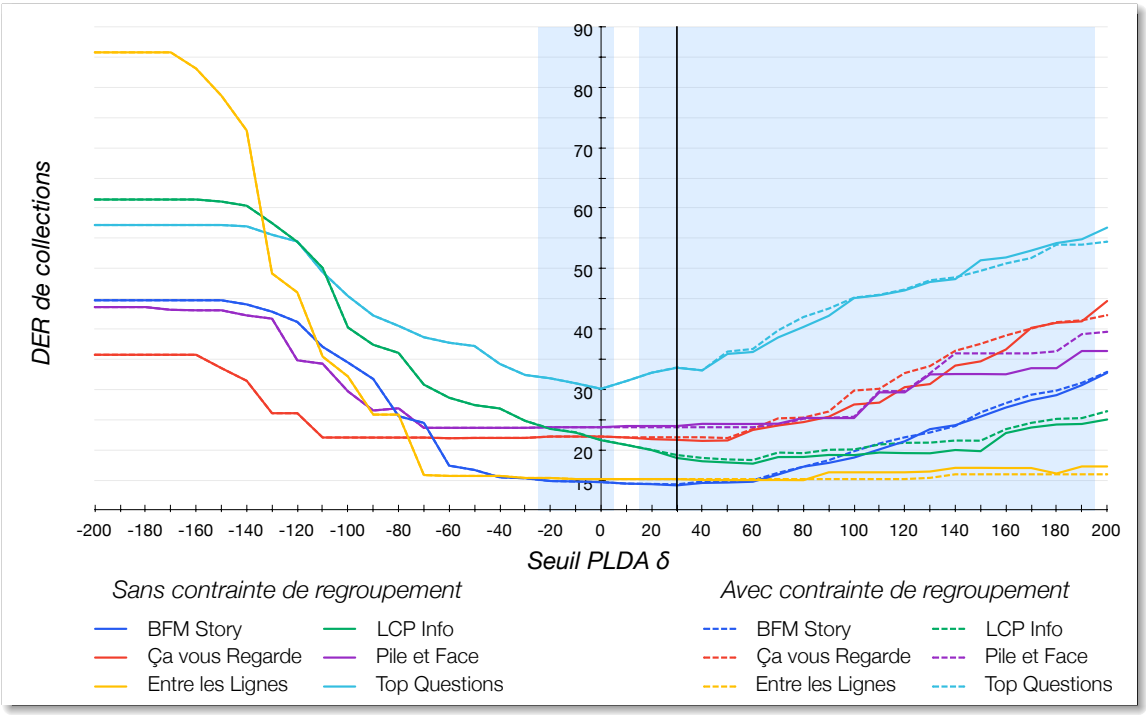

Figure 5.23 – **$DER_{\text{de collections}}$** *obtenus par regroupement global HAC_{PLDA}, pour différentes valeurs du seuil PLDA, pour les deux stratégies de regroupement global.*

Les aires rectangulaires sur la figure 5.23 représentent les zones où le $DER_{\text{de collections}}$ moyen obtenu avec la stratégie « sans contrainte » est inférieur à celui de la stratégie « avec contrainte ». Quelle que soit la collection étudiée, les résultats sont très proches entre les deux stratégies de regroupement, avec un $DER_{\text{de collections}}$ minimal pour le seuil $\delta = 30$, égal à 19,27% pour la stratégie « avec contrainte », et 19,08% pour la stratégie « sans contrainte ».

▷ Discussion

La différence entre les résultats obtenus par les deux stratégies de regroupement n'est pas aussi importante qu'attendu. Notre intuition quant à laisser la liberté au système de regrouper des classes issues d'un même enregistrement s'avère toutefois correcte au vu de la légère amélioration des $DER_{\text{d'émissions}}$ et $DER_{\text{de collections}}$ observée : certaines erreurs effectuées lors du regroupement du niveau *émission* de l'architecture de regroupement global, où les enregistrements sont traités séparément, sont corrigées par le regroupement global du niveau *collection*.

5.4. Analyse et bilan

Nous avons expérimenté notre architecture de regroupement global, dont la représentation schématique est rappelée en figure 5.24, sur les différentes collections constituées d'après les données REPERE et ETAPE (*cf.* chapitre 4). Cette partie se veut volontairement synthétique, elle constitue en quelque sorte un bilan général établi sur la base d'une analyse intermédiaire et approfondie des résultats expérimentaux obtenus sur chacune des collections étudiées. Cette analyse intermédiaire est présentée en détail en annexe A (p.175). À ce titre, certaines des observations énoncées dans cette partie sont redondantes avec celles présentées dans l'annexe A.

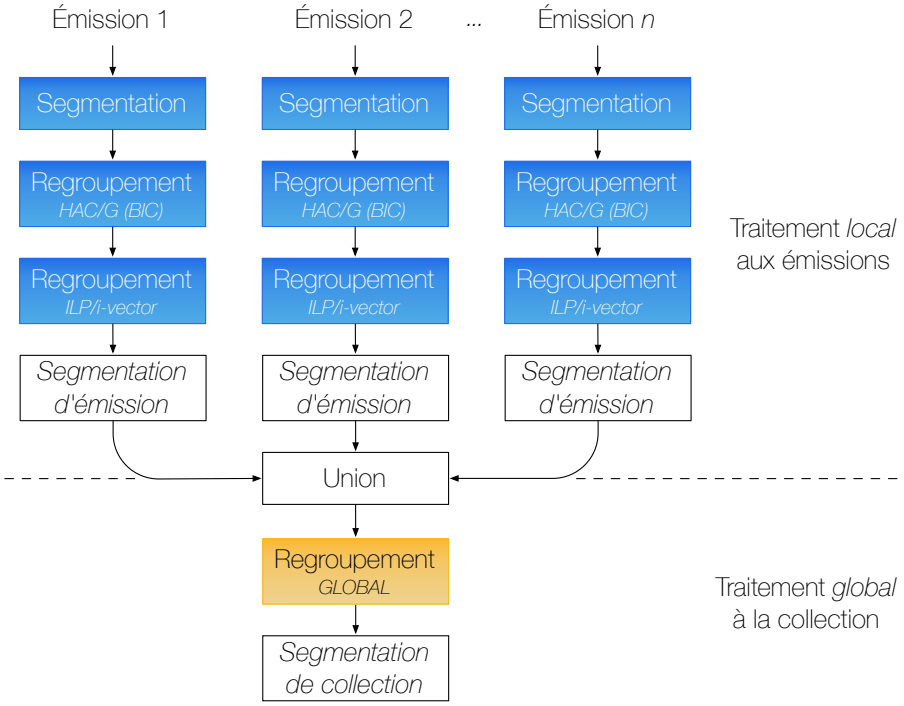

Figure *5.24 – Architecture de regroupement global pour la SRL de collections.*

Les approches expérimentées pour le regroupement global sont celles basées sur la décomposition en composantes connexes et les scores PLDA : le regroupement ILP ($CC{+}ILP_{PLDA}$) et le regroupement agglomératif hiérarchique ($CC{+}HAC_{PLDA}$). La stratégie adoptée pour effectuer le regroupement global consiste à laisser la liberté au système de regrouper des classes provenant d'un même enregistrement. La SRL d'émission, effectuée indépendamment sur chaque enregistrement, au niveau *émission* de l'architecture, correspond à celle présentée dans la section 5.2.2 : il s'agit d'un système de SRL d'émissions à l'état de l'art où les classes issues du regroupement BIC ($\lambda = 3$) sont modélisées par des i-vectors de dimension 300 et regroupées par l'approche ILP_{PLDA} avec un seuil $\delta = 20$.

Nous proposons dans un premier temps d'analyser les résultats recueillis sur les collections d'émissions de niveaux *programme*, *organisme (chaîne)* et *thématique (toutes les données)*. Nous présentons ensuite une analyse propre aux résultats expérimentaux obtenus avec les collections temporelles. Des observations d'ordre général seront présentées dans une troisième section.

5.4.1 Analyse sur les collections d'émissions

La composition des différentes collections d'émissions, déjà présentée dans le chapitre 4, est rappelée dans le tableau 5.13. Dans ce tableau, le nombre de locuteurs récurrents a été calculé à partir des segmentations de référence, il correspond au nombre de locuteurs intervenant dans au moins deux enregistrements différents.

Niveau	Collection		n^{bre} enr.	Durée		n^{bre} locuteurs	
				audio	UEM	total	récur.
Programme	BFMTV	**BFM Story** *(+ Ruth Elkrief)*	48	49h32	23h00	556	83
		Planète Showbiz *(+ Culture et Vous)*	160	38h27	5h00	771	75
	LCP	**Ça vous Regarde**	23	20h55	7h14	173	11
		Entre les Lignes	27	16h14	7h05	18	9
		LCP Info *(+ LCP Actu)*	48	20h34	11h14	317	96
		Pile et Face	33	19h44	6h11	46	15
		Top Questions	35	12h20	7h28	119	36
Organisme	**BFMTV**		208	88h00	28h00	1309	160
	LCP		166	89h48	39h13	544	172
Thématique	**BFMTV + LCP**		374	177h48	67h13	1787	333

Table *5.13 – Composition des collections d'émissions : nombre d'enregistrements, durée totale de la collection et durée évaluée (UEM), nombre de locuteurs total et récurrents*

Les seuils β, pour l'approche de décomposition en composantes connexes *(CC)*, et δ, pour les approches de regroupement ILP_{PLDA} et HAC_{PLDA} ont été choisi selon le principe énoncé dans la partie 5.2.3, afin de conserver une certaine cohérence lors de la comparaison des deux approches de regroupement : concernant l'approche de regroupement $CC{+}ILP_{PLDA}$, la valeur du seuil β doit être égale à celle du seuil δ, et concernant l'approche $CC{+}HAC_{PLDA}$, le seuil β doit être identique à celui déterminé pour l'approche $CC{+}ILP_{PLDA}$. Le critère d'optimisation pour la sélection de ces seuils, étant donné les deux « règles », repose sur le $DER_{\text{de collections}}$ moyen minimum. Les seuils ainsi retenus pour effectuer les expériences sur les collections d'émissions ne sont donc pas les mêmes selon le niveau de la collection d'émissions (*cf.* tableau 5.14).

Niveau	CC+ILP$_{PLDA}$		CC+HAC$_{PLDA}$	
	β	δ	β	δ
Programme	-30	-30	-30	30
Organisme	-20	-20	-20	60
Thématique	-20	-20	-20	60

Table *5.14 – Valeur des seuils β et δ sélectionnés pour les approches de regroupement* CC+ILP$_{PLDA}$ *et* CC+HAC$_{PLDA}$ *pour les différents niveaux de collections d'émissions.*

Nous présentons, en figure 5.25, les résultats obtenus sur les différentes collections d'émissions par les approches de regroupement global *CC+ILP$_{PLDA}$* et *CC+HAC$_{PLDA}$*, en termes de DER$_{\text{de collections}}$ et DER$_{\text{d'émissions}}$. Les deux approches de regroupement donnent des résultats très proches. Quelle que soit la collection d'émissions étudiée, la différence en DER$_{\text{de collections}}$ entre les deux approches de regroupement est inférieure à 2%, exception faite de la collection *Planète Showbiz* pour laquelle les résultats obtenus, avec les seuils β et δ sélectionnés sont très médiocres. Les DER$_{\text{d'émissions}}$ sont encore plus stables, avec une différence maximale de 0,1% entre les résultats des deux approches de regroupement.

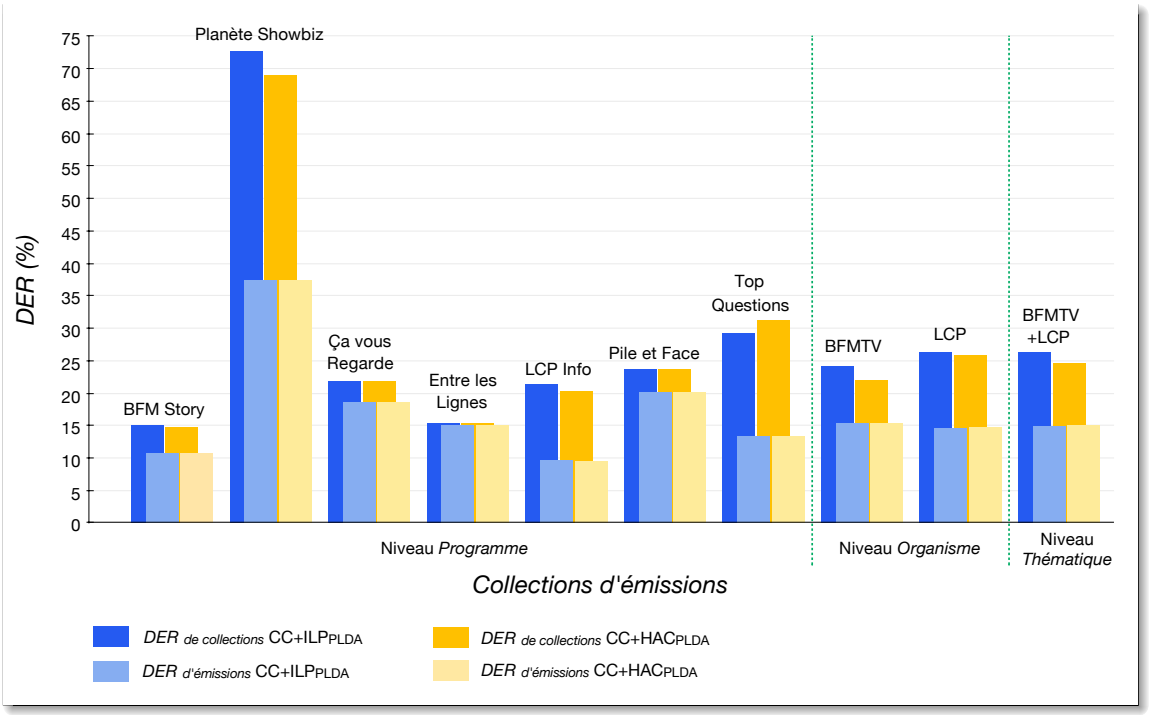

Figure *5.25 – Représentation des résultats obtenus par les approches de regroupement* CC+ILP$_{PLDA}$ *et* CC+HAC$_{PLDA}$*, en termes de* **DER$_{\text{de collections}}$** *et* **DER$_{\text{d'émissions}}$***, sur les différentes collections d'émissions.*

En moyenne, l'approche de regroupement *CC+HAC$_{PLDA}$* permet d'obtenir des DER$_{\text{de collections}}$ légèrement inférieurs à ceux obtenus par l'approche *CC+ILP$_{PLDA}$*, et plus la taille de la collection augmente, plus l'écart entre les résultats des deux ap-

proches de regroupement se creuse : cet écart, toujours en faveur du regroupement $CC{+}HAC_{PLDA}$ est de 0,11% au niveau *Programme*; 1,23% au niveau *Organisme*; 1,64% au niveau *Thématique*. Il est également intéressant de constater que la combinaison de seuils β et δ sélectionnée au niveaux *Organisme* et *Thématique* sont identiques. Il semblerait donc que plus la collection est volumineuse, plus la combinaison de seuils idéale se stabilise (pour les deux approches de regroupement). Il ne s'agit cependant que d'une conjecture, étant donné que la sélection de ces seuils a été réalisée en fonction des $\text{DER}_{\text{de collections}}$ moyens pour les niveaux *Programme* et *Organisme* des collections d'émissions.

Le $\text{DER}_{\text{de collections}}$ permet d'estimer la qualité des classifications produites en tenant compte des locuteurs récurrents, cependant, cette métrique d'évaluation ne permet pas d'estimer l'efficacité des approches de regroupement quant à la détection des locuteurs récurrents, car aucune distinction n'est faite entre locuteurs récurrents et non-récurrents. Toutefois, l'outil d'évaluation mis à notre disposition permet de générer la correspondance entre les étiquettes des segmentations de référence et celles produites par le système. Nous avons considéré qu'un locuteur récurrent est correctement détecté par le système si l'étiquette de la classe qui le représente est associée à des segments provenant d'au moins deux enregistrements différents, et bien sûr, si le locuteur correspondant dans les segmentations de référence est effectivement récurrent. Attention, la correspondance entre les locuteurs des segmentations de référence et celles produites par le système repose sur l'algorithme Hongrois [Galibert, 2013]. Notre conception d'un locuteur récurrent correctement détecté par le système ne garantit donc pas que locuteur ait été détecté dans tous les enregistrements où, d'après les références, il est censé intervenir.

Nous présentons, en figure 5.26, les résultats d'une analyse portant sur le nombre de locuteurs récurrents détectés par les approches de regroupement global. Sur cette figure, le nombre total de locuteurs récurrents détectés par les approches de regroupement global, pour chaque collection d'émissions, est représenté par une barre mi-claire, mi-foncée. La proportion de locuteurs récurrents correctement détectés (qui correspondent à des locuteurs récurrents d'après les segmentations de référence) est représentée en foncé. Les proportions de locuteurs récurrents sont exprimées en pourcentage du nombre de locuteurs récurrents présents dans les segmentations de référence, 100% (ligne rouge) étant l'objectif à atteindre.

Les approches de regroupement $CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$ détectent généralement moins de locuteurs récurrents qu'il n'y en a dans les segmentations de référence. En moyenne sur les collections d'émissions de niveau *Programme*, 67,9% des locuteurs récurrents détectés par l'approche de regroupement $CC{+}ILP_{PLDA}$ (respectivement, 69,6% pour l'approche $CC{+}HAC_{PLDA}$) correspondent effectivement à des

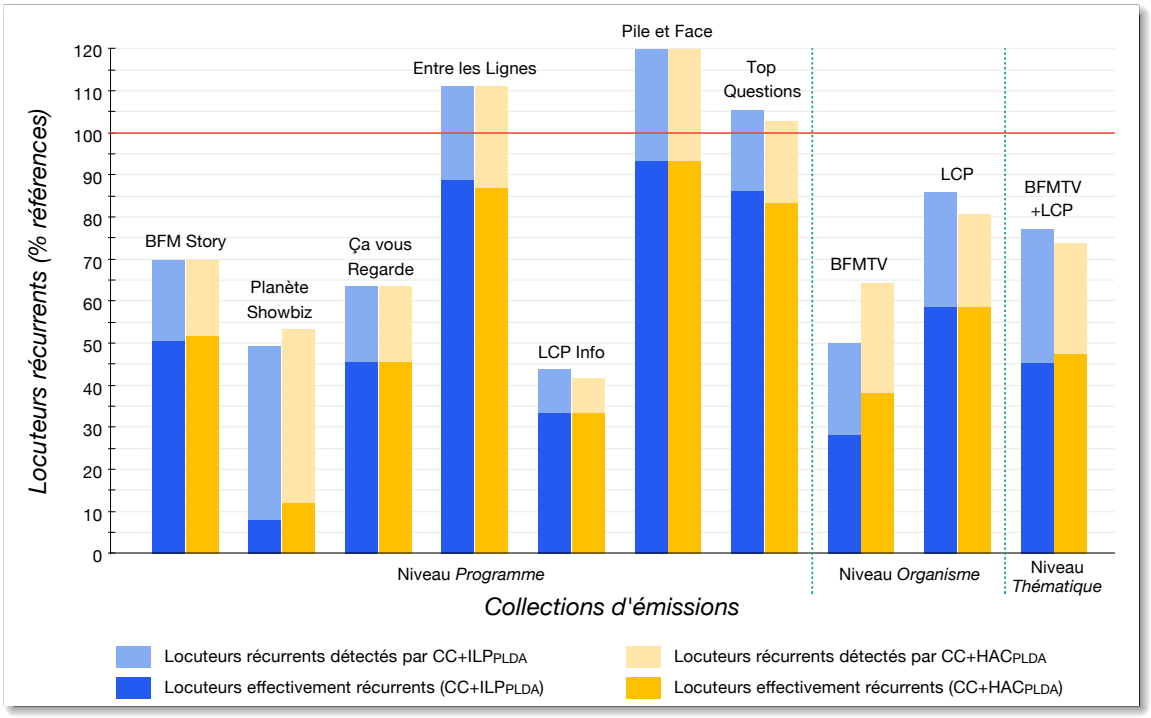

Figure *5.26 – Représentation du nombre de locuteurs récurrents détectés par les approches de regroupement* CC+ILP$_{\text{PLDA}}$ *et* CC+HAC$_{\text{PLDA}}$*, pour chacune des collections étudiées. La partie foncée des barres représente la proportion de locuteurs récurrents détectés par les approches de regroupement qui sont effectivement récurrents d'après les segmentations de référence.*

locuteurs récurrents d'après les segmentations de référence. La proportion de locuteurs récurrents correctement détectés diminue en fonction du volume des collections évaluées : cette proportion est de 62,3% *(CC+ILP$_{PLDA}$)* et 65,9% *(CC+HAC$_{PLDA}$)* sur les collections de niveau *Organisme*; elle chute à 58,8% *(CC+ILP$_{PLDA}$)* et 64,2% *(CC+HAC$_{PLDA}$)* sur la collection de niveau *Thématique*. Les proportions de locuteurs récurrents correctement détectés peuvent sembler faibles au regard des taux DER, où la différence entre DER$_{\text{de collections}}$ et DER$_{\text{d'émissions}}$ n'est que d'environ 10% en moyenne.

Il semble donc que plus le volume de la collection augmente, plus l'approche de regroupement *CC+ILP$_{PLDA}$* détecte, à tort, des locuteurs récurrents. Cette observation peut permettre de justifier l'écart observé entre les DER$_{\text{de collections}}$ des approches de regroupement *CC+ILP$_{PLDA}$* et*CC+HAC$_{PLDA}$*, qui est également fonction du volume de la collection.

Finalement, 58% des locuteurs récurrents ont été correctement détectés par les approches de regroupement *CC+ILP$_{PLDA}$* et *CC+HAC$_{PLDA}$*, en moyenne, sur les collections de niveau *Programme*. Sur les collections de niveau *Organisme*, l'approche *CC+ILP$_{PLDA}$* a permis de détecter correctement 43,4% des locuteurs récurrents, contre 48,4% pour l'approche *CC+HAC$_{PLDA}$*. Enfin, avec la collection de niveau

Thématique, ce sont 45,4% des locuteurs récurrents qui ont été correctement détectés par l'approche de regroupement $CC{+}ILP_{PLDA}$, contre 47,4% pour l'approche $CC{+}HAC_{PLDA}$.

5.4.2 Analyse sur les collections temporelles

La répartition des enregistrements en collections *temporelles* a été réalisée en fonction des « irrégularités » observées dans la fréquence d'acquisition des enregistrements. Cette répartition, présentée plus en détail dans le chapitre 4, est rappelée ci-dessous dans le tableau 5.15.

Collection	Période couverte	n^{bre} enr.	Durée		n^{bre} locuteurs
			audio	UEM	
Temporelle n°1	73 jours	11	6h11	3h52	[58 ; 10 ; **2**]
Temporelle n°2	30 jours	29	15h42	5h46	[217 ; 26 ; **9**]
Temporelle n°3	9 jours	20	10h20	2h45	[140 ; 28 ; **14**]
Temporelle n°4	105 jours	114	39h44	21h01	[669 ; 108 ; **42**]
Temporelle n°5	51 jours	46	18h02	6h26	[247 ; 51 ; **17**]
Temporelle n°6	44 jours	32	9h29	4h19	[205 ; 30 ; **0**]
Temporelle n°7	26 jours	24	12h36	8h38	[262 ; 37 ; **0**]
Temporelle n°8	121 jours	90	61h55	13h21	[461 ; 64 ; **27**]

Table *5.15 – Composition des collections temporelles : période couverte en nombre de jours, nombre d'enregistrements, durée totale et durée évaluée (UEM), nombre de locuteurs (formalisme [n^{bre} locuteurs total ; n^{bre} locuteurs récurrents ; n^{bre} locuteurs récurrents sur des enregistrements provenant d'émissions différentes]).*

Les seuils β et δ ont été sélectionnés de manière à minimiser le $DER_{\text{de collections}}$ moyen sur les 8 collections temporelles. Pour l'approche de regroupement global $CC{+}ILP_{PLDA}$, $\beta = \delta = 10$, et pour l'approche de regroupement $CC{+}HAC_{PLDA}$, $\beta = 10$, et $\delta = 40$.

Les deux approches de regroupement global donnent des résultats similaires, comme en témoigne la figure 5.27. On n'observe qu'une très faible différence entre les $DER_{\text{de collections}}$ des collections traitées, avec en moyenne 20,66% pour l'approche $CC{+}ILP_{PLDA}$ et 20,37% pour l'approche $CC{+}HAC_{PLDA}$. Ces résultats moyens sont difficilement comparables avec ceux obtenus sur les collections d'émissions, étant donné la composition hétérogène des collections temporelles. En revanche, on notera que les $DER_{\text{de collections}}$ moyens obtenus sont sensiblement inférieurs à ceux des collections d'émissions, avec un gain absolu de 2,63% pour l'approche de regroupement $CC{+}ILP_{PLDA}$ (respectivement, 2,81% pour l'approche $CC{+}HAC_{PLDA}$) par rapport aux *meilleurs* $DER_{\text{de collections}}$ des collections d'émissions (obtenus au niveau *Programme*).

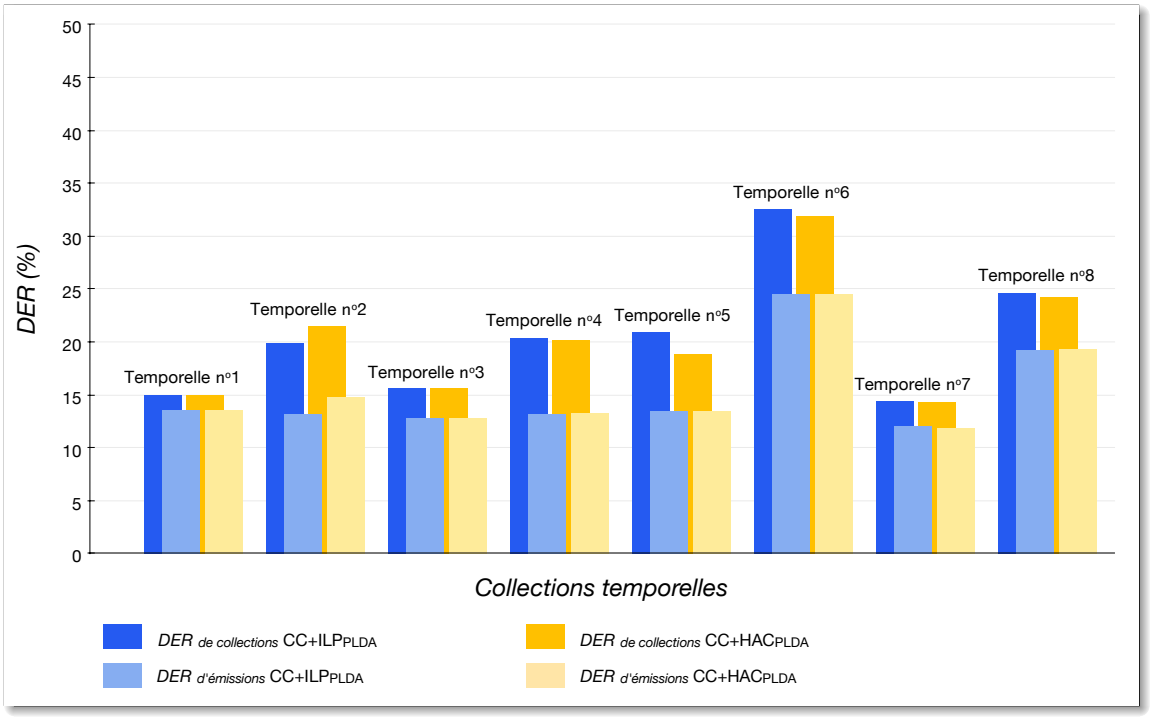

Figure *5.27 – Représentation des résultats obtenus par les approches de regroupement* CC+ILP$_{\text{PLDA}}$ *et* CC+HAC$_{\text{PLDA}}$*, en termes de* **DER$_{\text{de collections}}$** *et* **DER$_{\text{d'émissions}}$**, *sur les 8 collections temporelles.*

Une représentation synthétique de l'analyse sur les locuteurs récurrents est présentée en figure 5.28. En moyenne, 70% des locuteurs récurrents détectés par l'approche de regroupement *CC+ILP$_{PLDA}$* sont effectivement récurrents d'après les segmentations de référence. Avec l'approche de regroupement *CC+HAC$_{PLDA}$*, cette proportion est de 69,8%, donc presque identique. Sur l'ensemble des collections temporelles, les deux approches de regroupement global détectent correctement 65% des locuteurs récurrents, en moyenne. Il s'agit du meilleur taux de détection en locuteurs récurrents obtenu avec cette architecture de regroupement global. Cette particularité peut s'expliquer par le fait que les collections temporelles sont composées d'enregistrements de provenance diverse, et les seuils β et δ, qui ont été sélectionnés pour minimiser le DER$_{\text{de collections}}$, sont donc probablement plus robustes. Comparé au meilleur taux de détection en locuteurs récurrents sur les collections d'émissions (58%, pour les collections d'émissions de niveau *Programme*), le gain absolu est de 7%.

Le découpage en collection temporelle, où les enregistrements des différentes émissions à notre disposition sont mélangés, semble donner des résultats supérieurs à ceux obtenus avec les collections d'émissions : les résultats en termes de DER$_{\text{de collections}}$ sont inférieurs, et la proportion de locuteurs récurrents détectés est supérieure. Il est cependant difficile de tirer une conclusion à ce sujet, car la plus volumineuse des collections temporelles ne représente qu'une quarantaine d'heures de données audio. De plus, le nombre de locuteurs récurrents dans les collections tem-

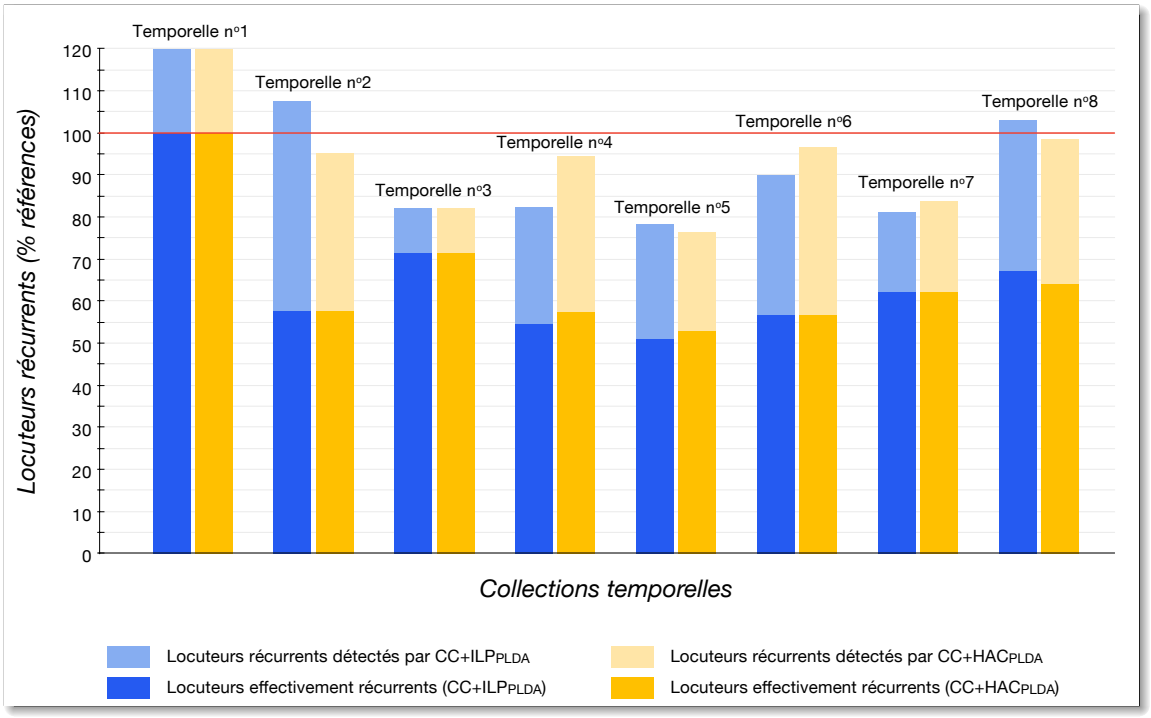

Figure *5.28 – Représentation du nombre de locuteurs récurrents détectés par les approches de regroupement* CC+ILP$_{\text{PLDA}}$ *et* CC+HAC$_{\text{PLDA}}$*, pour chacune des collections temporelles. La partie foncée des barres représente la proportion de locuteurs récurrents détectés par les approches de regroupement qui sont effectivement récurrents d'après les segmentations de référence.*

porelles est inférieur au nombre de locuteurs récurrents présents dans les collections d'émissions.

5.4.3 Observations générales

L'architecture de regroupement global, où l'ensemble des enregistrements composant une collection est traité simultanément, est efficace pour traiter les collections volumineuses à notre disposition dans la mesure où les approches de regroupement mises en œuvre sont capables de gérer un grand nombre de classes dans un laps de temps raisonnable. Les taux d'erreurs DER$_{\text{de collections}}$, qui globalement augmentent en fonction de la taille de la collection, restent toutefois acceptables au regard des taux DER$_{\text{d'émissions}}$ qui, eux, restent stables et relativement faibles compte tenu de la diversité acoustique des données étudiées.

Les approches de regroupement ILP et HAC reposant sur l'approche de simplification par décomposition en composantes connexes sont très rapides, quelle que soit la collection étudiée. En revanche, la durée nécessaire à l'estimation des mesures de similarité (scores PLDA) entre les classes impliquées dans les regroupements est beaucoup plus importante, comme en témoigne le graphique présenté en figure 5.29

(échelle logarithmique sur l'axe des abscisses).

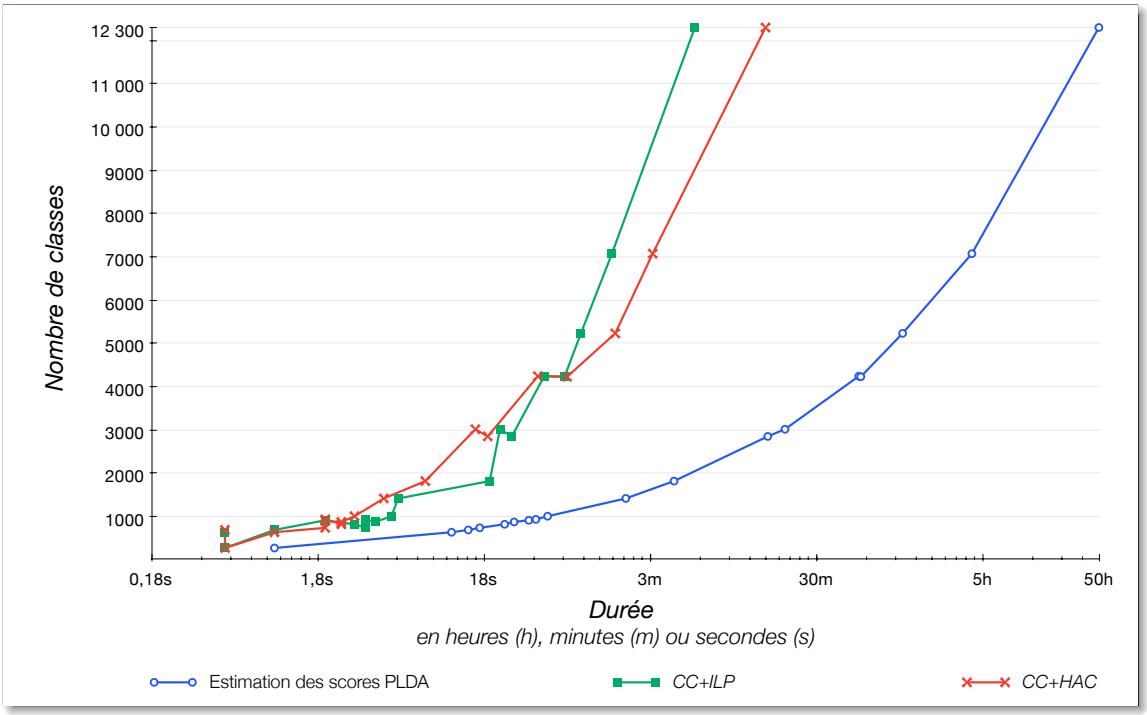

Figure *5.29 – Évolution des durées de traitement en fonction du nombre de classes impliquées dans le problème de regroupement (architecture de regroupement global). Les données ont été recueillies sur chacune des collections traitées (échelle logarithmique sur l'axe des abscisses).*

Le calcul des scores PLDA est réalisé avec l'outil *IvTest* de la suite d'outils pour la reconnaissance du locuteur *Alize* [Bonastre et al., 2008]. L'outil *IvTest* a été configuré pour répartir la charge de calcul sur quatre fils d'exécution (threads). Les durées présentées quant à l'estimation des scores PLDA, qui correspondent au temps d'exécution de l'outil *IvTest*, pourraient toutefois être amoindries en augmentant le nombre de fils d'exécution (à condition de disposer de processeurs en quantité suffisante). Ces durées pourraient également être diminuées en considérant la propriété de symétrie des distances (étant donné deux modèles de locuteurs i et j, $S_{PLDA}(i,j) = S_{PLDA}(j,i)$). Ainsi, par rapport à la collection de niveau *Thématique* dont la durée effective du calcul des scores PLDA est d'environ 50h, nous estimons grossièrement (et de manière optimiste) que cette durée pourrait être divisée par 10 en considérant la propriété de symétrie des distances (durée totale divisée par 2), et 20 fils d'exécution au lieu de 4 (durée totale divisée par 5). Cette durée hypothétique (5 heures) reste néanmoins considérable comparé aux durées des approches de regroupement *CC+ILP* (11 minutes) et *CC+HAC* (30 minutes). Nous n'avons malheureusement pas pris conscience de l'importance de ces optimisations avant de traiter nos collections les plus volumineuses (niveau *Organisme* et *Thématique*), car la durée nécessaire à l'estimation des scores PLDA nous a semblé raisonnable pour les

collections dont le nombre de classes impliquées dans les problèmes de regroupement est inférieur à 5000.

L'approche de regroupement *CC+ILP* donne des résultats légèrement moins bons, en termes de $\text{DER}_{\text{de collections}}$, quelle que soit la collection étudiée. Le procédé de résolution du problème ILP échoue si le problème est trop complexe. Il est possible qu'un outil de résolution différent, plus récent et plus performant, permette de s'affranchir des problèmes rencontrés quant à la complexité des problèmes soumis par rapport au seuil δ. Durant la première année de cette thèse, nous avions expérimenté l'outil de résolution dénommé Gurobi [Gurobi Optimization, Inc., 2015] qui, s'il n'était peut-être pas plus performant, semblait plus rapide pour estimer les solutions. Nous avons cependant préféré continuer à employer l'outil distribué par la fondation GNU en raison de certaines contraintes d'ordre technique (essentiellement dues à la licence académique proposée par Gurobi), et juridique (l'intégration de cet outil au *LIUM_spkDiarization Toolkit* n'étant pas techniquement possible avec la licence académique).

En comparaison, l'approche de regroupement *CC+HAC* est plus robuste (le temps de calcul est toujours resté en dessous de la contrainte fixée). Les $\text{DER}_{\text{de collections}}$ obtenus sont légèrement inférieurs à ceux obtenus avec l'approche *CC+ILP*, et l'écart entre les résultats des deux approches de regroupement semble se creuser en fonction du volume de données à traiter. Les $\text{DER}_{\text{d'émissions}}$ sont, en revanche, très légèrement supérieurs à ceux obtenus avec l'approche *CC+ILP*. Cette différence est cependant presque imperceptible, de l'ordre de 0.1% dans le pire des cas avec les collections étudiées. En termes de durées, pour ce qui est de ces approches de regroupement, on constate que plus le nombre total de classes impliquées dans les problèmes de regroupement augmente, plus l'approche *CC+ILP* se démarque de l'approche *CC+HAC* en rapidité d'exécution, en particulier pour traiter des problèmes constitués de plus 5000 classes. Cette observation peut sembler étrange compte tenu du fait que l'approche de classification ILP n'est pas exécutée en temps polynomial, contrairement à HAC. Toutefois, les deux approches de classification ne sont pas réalisées par le même outil, des différences dans leurs implémentations respectives peuvent expliquer l'écart observé en termes de rapidité d'exécution. La durée d'exécution des approches de regroupement reste cependant négligeable au regard du temps nécessaire à l'estimation des scores de vraisemblance.

CHAPITRE 6

SRL de collections par regroupement incrémental

Le chapitre précédent a présenté nos travaux sur le regroupement global, où les différents enregistrements constituant une collection sont traités simultanément. Nous présentons, dans ce chapitre, nos travaux sur le regroupement incrémental, où les enregistrements sont traités les uns à la suite des autres. L'architecture que nous présentons, également inspirée par les travaux antérieurs menés par [Rouvier et Meignier, 2012; Tran et al., 2011; Yang et al., 2011], s'intègre davantage dans le cadre du projet européen EUMSSI [1] pour lequel le LIUM, en charge de la modalité audio, doit traiter de très grandes quantités de données.

Nous rappelons dans une première partie l'intérêt d'une architecture par regroupement incrémental par rapport à son alternative par regroupement global, en discutant les avantages, les inconvénients et les limites propres aux deux approches. Nous décrivons ensuite l'architecture par regroupement incrémental mise en œuvre et la configuration employée. Les parties suivantes présentent des spécificités expérimentées dans le but de réduire à la fois la complexité du regroupement incrémental, et les taux d'erreur $\text{DER}_{\text{de collections}}$. Dans une dernière partie, nous évaluons notre approche incrémentale sur les collections d'émission et les collections temporelles construites à partir des données ETAPE et REPERE (*cf.* chapitre 4), de manière à établir et discuter les performances de notre approche en comparaison aux résultats obtenus avec l'architecture de regroupement global.

1. Event Understanding through Multimodal Social Stream Interpretation

6.1. Contexte et approche envisagée

Le regroupement incrémental (ou *séquentiel*) de différents enregistrements est un concept abordé par Leeuwen [2010] dans le cadre de l'appariement en locuteurs *on-line*. Dans [Tran et al., 2011; Yang et al., 2011], les auteurs proposent une architecture de regroupement incrémental orientée collection, dans laquelle un module d'identification en locuteur (*Open-Set Identification* – OSI) est utilisé pour reconnaître les locuteurs récurrents déjà détectés dans les enregistrements préalablement traités. Dans cette partie, nous détaillons dans un premier temps les avantages et inconvénients de cette approche de regroupement, déjà mentionnés dans le chapitre 3. Nous présentons ensuite notre architecture de regroupement incrémental.

6.1.1 Spécificités et limites

Il a été établi par Tran et al. [2011]; Yang et al. [2011] que l'approche de regroupement incrémental présente deux inconvénients, en comparaison à l'approche de regroupement global : les résultats obtenus en termes de $\text{DER}_{\text{de collections}}$ sont plus élevés, et l'ordre dans lequel les enregistrements sont traités affecte ces résultats. En effet, l'approche de regroupement incrémental présente les mêmes caractéristiques que la méthode de classification hiérarchique : les erreurs de regroupement se propagent au fil des itérations. Ce phénomène est d'autant plus accentué que l'architecture présentée par les auteurs repose sur une approche de regroupement hiérarchique. Il y a donc deux niveaux de propagation d'erreurs :

1. Le premier est local aux enregistrements traités. À chaque itération, des erreurs de regroupement irréversibles sont effectuées par l'algorithme de regroupement agglomératif hiérarchique (*cf.* partie 2.3.3).
2. Le deuxième niveau de propagation d'erreur, propre à l'ensemble de la collection, est dû à l'aspect itératif de l'architecture et son module d'identification du locuteur. Le module OSI permet de détecter les locuteurs déjà vus dans les enregistrements préalablement traités. Les modèles de locuteur présents dans le module OSI sont mis à jour à l'issue de chaque itération i avec les données provenant de l'enregistrement traité durant l'itération i. L'inconvénient de cette approche est que la robustesse des modèles de locuteur présents dans le modèle OSI est fonction du nombre d'itérations effectuées : plus les itérations passent et plus les modèles contenus dans le module OSI deviennent robustes. Les locuteurs récurrents des premiers enregistrements traités ne sont donc pas détectés avec le même degré d'objectivité que dans les derniers enregistrements.

Il n'est donc pas surprenant que les taux d'erreur soient supérieurs à ceux obtenus par une approche de regroupement global, étant donné que le système dispose d'une vision d'ensemble du problème de regroupement, et que les performances de la détection des locuteurs récurrents dépendent avant tout de l'ordre dans lequel les émissions sont traitées. Les résultats en termes de $DER_{\text{de collections}}$ obtenus par [Tran et al., 2011] avec l'approche de regroupement incrémental sont compris entre 17,8% et 20,8%, selon l'ordre dans lequel sont traités les enregistrements, alors que l'approche de regroupement global permet d'atteindre un $DER_{\text{de collection}}$ de 15,2%.

L'approche de regroupement incrémental présente néanmoins un attrait de taille pour traiter des collections volumineuses, et dans une certaine mesure, des collections dont la taille augmenterait dynamiquement en fonction du temps. L'architecture que nous présentons dans la partie suivante, qui est toujours en cours d'investigation dans le projet de recherche européen EUMSSI, tire parti des travaux réalisés sur le regroupement global, en particulier, la classification par les graphes (décomposition en composantes connexes).

6.1.2 Architecture proposée pour le regroupement incrémental des collections

Nous proposons une architecture de regroupement incrémental où les différents enregistrements qui composent une collection seraient traités itérativement selon un critère d'ordre correspondant à la date de diffusion des enregistrements (*cf.* figure 6.1). Ce type d'architecture est propice au traitement des collections dont la taille augmenterait dynamiquement en fonction du temps, contrairement à l'approche de regroupement global où il serait nécessaire de recommencer l'intégralité du procédé de SRL de collections si de nouveaux enregistrements venaient enrichir une collection existante déjà traitée. Il est très fréquent, avec les émissions journalistiques d'information, qu'un même sujet d'actualité soit traité sur plusieurs jours d'affilés au moment des faits, afin d'en couvrir l'évolution générale. Il est possible, dans cette situation, que des locuteurs jusqu'alors *anonymes* viennent à se faire connaître pour leur affinité avec l'évènement relaté. Si ces locuteurs anonymes sont fréquemment interrogés au cours de la médiatisation d'un évènement, ils deviennent alors récurrents. La question qui se pose est comment détecter de tels locuteurs récurrents dans une collection en constante évolution ? Prenons le cas d'une collection d'émissions de niveau *Programme* comme *LCP Info*, qui relate l'actualité au quotidien, que nous compléterions chaque jour par les nouveaux enregistrements obtenus. Nous serions alors probablement confrontés à des enregistrements présentant en tout ou partie l'évolution de la situation pour un sujet d'actualité donné. Un acteur *anonyme* ayant

un lien avec le sujet traité pourrait tout d'abord être cité pour son implication. À mesure que la situation évolue, cet acteur *anonyme* pourrait être interviewé par un journaliste, ou être invité à participer à l'émission. Selon son implication, son avis pourrait même être sollicité à plusieurs reprises, au même titre qu'un expert. Le problème dans ce contexte, c'est qu'au jour d'aujourd'hui nous ne pouvons pas anticiper la récurrence d'un locuteur. S'il est acceptable de ne pas, ou mal, détecter un locuteur dans un enregistrement isolé, c'est beaucoup plus problématique si le locuteur intervient dans plusieurs enregistrements. Or le regroupement incrémental des collections présente un inconvénient : les regroupements effectués lors d'une itération sont définitifs, les erreurs ne peuvent pas être corrigées. Nous pensons que respecter l'ordre chronologique des enregistrements revient, dans une certaine mesure, à respecter la chronologie des évènements relatés.

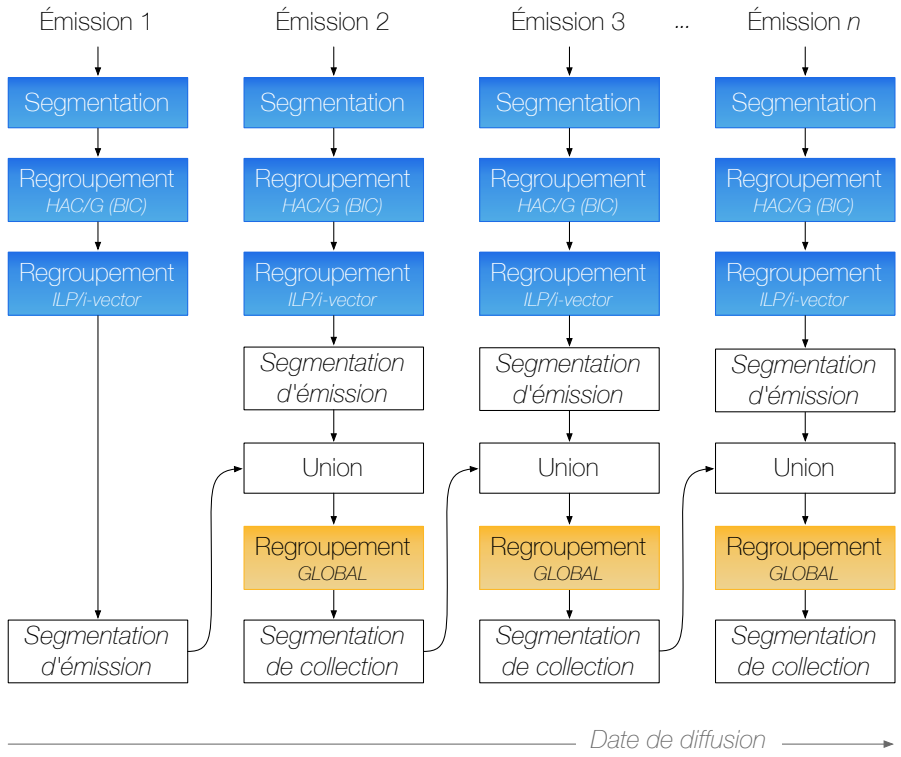

Figure *6.1 – Architecture par regroupement incrémental pour la SRL de collections.*

Le principe de l'architecture de regroupement incrémental que nous proposons, où les enregistrements sont traités itérativement en fonction de l'ordre chronologique de diffusion, est illustré en figure 6.1. Cette architecture de regroupement est dite « séquentielle » dans la mesure où les enregistrements sont traités les uns à la suite des autres, par opposition au regroupement global qui traite l'ensemble des enregistrements simultanément.

Le principe est le suivant : à chaque itération i de l'algorithme de regroupement,

l'enregistrement courant est d'abord traité par le système de SRL d'émissions habituel, délivrant une segmentation d'émission où les n locuteurs de cet enregistrement sont censés être *idéalement* répartis en n classes distinctes. La segmentation ainsi obtenue est ensuite concaténée à la segmentation de collections courante, résultat de l'itération $i-1$, dans laquelle les locuteurs récurrents des enregistrements précédemment traités ont été *idéalement* détectés et regroupés au sein d'une classe unique à la collection. S'ensuit alors un regroupement global sur la concaténation des segmentations. L'objectif est de regrouper les classes de locuteur provenant de l'enregistrement courant avec celles correspondant aux locuteurs de la segmentation de collections courante. S'il y a regroupement, c'est que la classe représentant un locuteur de l'enregistrement courant est suffisamment proche d'une classe issue de la segmentation de collections, et donc, que le locuteur correspondant a déjà été détecté dans un enregistrement précédent. Le traitement du tout premier enregistrement de la collection est particulier, puisqu'aucun regroupement global ne peut être effectué ne consiste qu'en déterminer sa segmentation d'émission. Nous présentons dans la partie suivante les résultats obtenus avec cette architecture sur les sept collections de niveau *Programme*.

6.1.3 Expérimentation

Les segmentations d'émissions des différents enregistrements composant nos collections d'émissions sont les mêmes que celles utilisées pour expérimenter l'architecture de regroupement global. Pour rappel, ces segmentations d'émissions ont été obtenues avec un système de SRL d'émission similaire à celui présenté dans la partie état de l'art si ce n'est que la dernière étape de regroupement correspond à un regroupement ILP_{PLDA} utilisant, en entrée, des classes de locuteurs déterminées par le regroupement BIC. Celles-ci ont été modélisées par des modèles i-vector de dimension 300 (extraits avec un GMM-UBM de 1024 composantes gaussiennes). Le seuil δ, correspondant à la valeur du score PLDA à partir de laquelle un regroupement entre deux classes n'est plus toléré, avait été fixé à 20. En moyenne sur l'ensemble des enregistrements des sept collections étudiées, le taux d'erreur $\text{DER}_{\text{d'émissions}}$ obtenu est de 13,17% et le taux d'erreur $\text{DER}_{\text{de collection}}$ est de 54,83% (les locuteurs récurrents ne sont pas détectés par le système de SRL d'émission).

Les seuils δ et β utilisés pour effectuer le regroupement global des différentes itérations sont ceux déterminés dans le chapitre précédent. Il s'agit des seuils ayant permis d'atteindre les meilleurs taux d'erreur $\text{DER}_{\text{de collections}}$ avec l'architecture de regroupement global. Nous présentons les résultats obtenus sur les sept collections du niveau *programme*, avec les approches de regroupement global *CC+ILP*$_{PLDA}$ et

$CC{+}HAC_{PLDA}$, dans les tableaux 6.1 ($DER_{\text{de collections}}$) et 6.2 ($DER_{\text{d'émissions}}$). Ces résultats sont comparés à ceux obtenus avec l'architecture de regroupement global (les valeurs entre parenthèses correspondent aux résultats obtenus avec l'approche de regroupement global).

Configuration	ILP_{PLDA} niveau émission	CC. + ILP_{PLDA} niveau collection	CC. + HAC_{PLDA} niveau collection
Seuil β	-	-30	-30
Seuil δ	20	-30	30
BFM Story	44,75%	14,32% *(15,03%)*	15,65% *(14,72%)*
Planète Showbiz	78,76%	47,35% *(72,74%)*	47,35% *(69,06%)*
Ça vous Regarde	35,73%	21,90% *(21,82%)*	21,90% *(21,82%)*
Entre les Lignes	85,77%	14,58% *(15,34%)*	14,58% *(15,34%)*
LCP Info	61,42%	20,86% *(21,30%)*	20,89% *(20,20%)*
Pile et Face	43,62%	23,41% *(23,74%)*	23,41% *(23,74%)*
Top Questions	57,20%	28,23% *(29,24%)*	28,23% *(31,28%)*
Moyenne	**54,83%**	**20,92% (23,29%)**	**21,38% (23,18%)**

Table *6.1* – **$DER_{\text{de collections}}$** *obtenus sur les collections de niveau* programme *avec l'architecture de regroupement incrémental pour les approches de regroupement $CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$. Les résultats entre parenthèses correspondent aux $DER_{\text{de collections}}$ obtenus sur les mêmes collections, avec les mêmes seuils, par l'architecture de regroupement global.*

Les $DER_{\text{de collections}}$ obtenus sur les sept collections d'émissions de niveau *Programme* avec l'architecture de regroupement incrémental sont globalement meilleurs que ceux obtenus avec l'architecture de regroupement global : la méthode de classification $CC{+}ILP_{PLDA}$ permet d'atteindre 20,92% (contre 23,29% pour la même méthode avec l'approche de regroupement global). La méthode $CC{+}HAC_{PLDA}$ permet quant à elle d'atteindre 21,38% (contre 23,18% avec l'approche de regroupement global). Ces premiers résultats sont très surprenants. Nous nous attendions, d'après les observations et résultats présentés par Tran et al. [2011], à obtenir des $DER_{\text{de collections}}$ supérieurs à ceux obtenus par l'approche de regroupement global. Dans le détail, la plupart des résultats obtenus avec l'approche incrémentale sont très proches, et souvent inférieurs, à ceux obtenus avec l'approche de regroupement global. Le principal gain est relatif à la collection *Planète Showbiz*, pour laquelle l'approche de regroupement global donnait des résultats très médiocres (les seuils β et δ sont pourtant identiques).

Les $DER_{\text{d'émissions}}$ restent relativement stables, comparé à ceux obtenus avec l'approche de regroupement global. Les approches $CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$ pour le regroupement incrémental se comportent de la même manière que pour le regroupement global : l'approche $CC{+}ILP_{PLDA}$ permet de maintenir (dans le cas présent, améliore légèrement) le $DER_{\text{d'émissions}}$ par rapport au $DER_{\text{d'émissions}}$ moyen obtenu sur les segmentations d'émissions (14,79% contre 14,92%). L'approche

Configuration	ILP_{PLDA} niveau émission	CC. + ILP_{PLDA} niveau collection	CC. + HAC_{PLDA} niveau collection
Seuil β	-	-30	-30
Seuil δ	20	-30	30
BFM Story	10,70%	10,35% *(10,70%)*	11,09% *(10,70%)*
Planète Showbiz	37,41%	36,49% *(37,41%)*	36,49% *(37,41%)*
Ça vous Regarde	18,51%	18,51% *(18,51%)*	18,51% *(18,51%)*
Entre les Lignes	15,01%	14,24% *(15,01%)*	14,24% *(15,01%)*
LCP Info	9,60%	10,05% *(9,60%)*	10,17% *(9,51%)*
Pile et Face	20,05%	19,72% *(20,05%)*	19,72% *(20,05%)*
Top Questions	13,31%	14,24% *(13,31%)*	14,24% *(13,31%)*
Moyenne	**14,92%**	**14,79% (14,92%)**	**15,06% (14,97%)**

Table *6.2* – **$DER_{d'émissions}$** *obtenus sur les collections de niveau* programme *avec l'architecture de regroupement incrémental pour les approches de regroupement CC+ILP_{PLDA} et CC+HAC_{PLDA}. Les résultats entre parenthèses correspondent aux $DER_{d'émissions}$ obtenus sur les mêmes collections, avec les mêmes seuils, par l'architecture de regroupement global.*

CC+HAC_{PLDA} donne quant à elle des $DER_{d'émissions}$ très proches, mais quelque peu supérieurs :+0,09% par rapport au $DER_{d'émissions}$ moyen obtenu avec l'architecture de regroupement global, et +0,14% par rapport au $DER_{d'émissions}$ moyen obtenu avec les segmentations d'émissions.

Il semble, au vu de ces premiers résultats sur les collections d'émissions du niveau *Programme*, que l'approche de regroupement *CC+ILP_{PLDA}* permette d'atteindre de meilleurs résultats que l'approche de regroupement *CC+HAC_{PLDA}*. Ce constat peut sembler intrigant dans la mesure où, avec l'architecture de regroupement global, nous avons observé le comportement contraire. Le comportement des méthodes de classification *CC+HAC_{PLDA}* et *CC+ILP_{PLDA}* sont cependant difficilement comparables entre les approches de regroupement global et incrémental : avec l'approche de regroupement incrémental, de nouveaux modèles de locuteurs sont calculés lors d'une itération n pour représenter les classes regroupées durant l'itération $n-1$, contrairement à l'approche de regroupement global.

6.1.4 Discussion

L'inconvénient le plus préjudiciable est l'absence de vision globale du problème de regroupement sur la collection. En effet, pour une collection constituée de n enregistrements, $n-1$ regroupements globaux seront itérativement effectués, provoquant des erreurs irrattrapables. L'architecture par regroupement global n'effectuera quant à elle qu'un seul regroupement sur l'ensemble des classes des n enregistrements, disposant ainsi du maximum d'information pour décider de la meilleure stratégie de

regroupement. Toutefois, bien qu'une multitude de regroupements soient effectués par l'architecture de regroupement incrémental, leur complexité est moindre, car moins de classes candidates sont impliquées. Bien sûr, la taille du problème de regroupement croît d'itération en itération, étant donné les concaténations itératives. Cependant, le seuil de décision en dessous duquel un regroupement entre deux classes est autorisé étant invariable au cours des itérations, les classes qui n'ont pas été regroupées durant une itération ne le seront pas plus dans les itérations suivantes. Donc finalement, pour une itération i donnée, les couples de classes impliqués dans le regroupement sont uniquement formés par les classes provenant de la segmentation d'émission de l'enregistrement traité lors de l'itération i, et les classes de la segmentation de collection. De plus, des regroupements sont effectués à chaque itération. Le regroupement global de l'itération correspondant au traitement du dernier enregistrement impliquera moins de classes que le regroupement global effectué par l'architecture de regroupement global.

6.2. Recyclage des modèles de locuteur

L'architecture de regroupement incrémental pour le traitement des collections volumineuses présente deux inconvénients : d'une part, les erreurs se propagent d'itération en itération, et d'autre part, le procédé est long comparé à l'approche de regroupement global. Nous avons obtenu des résultats positifs avec notre implémentation du regroupement incrémental (les résultats en termes de $\text{DER}_{\text{de collections}}$ sont inférieurs à ceux obtenus avec l'architecture de regroupement global). L'inconvénient relatif à la propagation des erreurs ne semble donc pas trop préjudiciable. En revanche, l'inconvénient lié à la durée du procédé de regroupement incrémental est plus gênant. Afin de minimiser ce problème, nous proposons une approche incrémentale alternative où, plutôt que d'extraire un nouveau modèle de locuteur pour représenter une nouvelle classe issue d'un regroupement, nous réutilisons le modèle de locuteur de l'une des classes ayant été regroupées.

Si des classes ont été regroupées lors d'une itération, c'est qu'*a priori* les données acoustiques représentées par ces classes proviennent d'un même locuteur. Les données acoustiques des classes regroupées sont habituellement utilisées pour apprendre un nouveau modèle de locuteur, censé mieux représenter la nouvelle classe obtenue. Ainsi, lors de chaque itération i de notre approche de regroupement incrémental, de nouveaux modèles de locuteurs sont extraits pour les classes correspondant à des regroupements effectués lors de l'itération $i-1$. Cette approche communément employée peut cependant être remise en question dans le contexte des architectures

de SRL de collections actuelles. En effet, cette approche revient à considérer que plus la quantité de données acoustiques disponible pour modéliser un locuteur est importante, plus le modèle correspondant est robuste. Il s'agit d'ailleurs du concept suivi pour mener à bien la dernière étape de regroupement en SRL d'émissions, où les classes issues du regroupement BIC, représentées par des modèles mono-gaussiens, sont suffisamment pures et en quantité suffisante pour permettre l'apprentissage des modèles GMM.

Jusqu'à présent, nous avons suivi la même recette pour effectuer les regroupements au niveau *collection* des architectures. Nous avons considéré que les regroupements effectués au niveau *émission* étaient suffisamment corrects pour extraire de nouveaux modèles de locuteurs étant donné les classes regroupées. Or, il est tout à fait concevable d'envisager qu'à ce niveau, utiliser l'ensemble des données acoustiques correspondant aux classes regroupées pourrait produire des modèles de locuteurs moins spécifiques, à cause de la variabilité indésirable introduite par les erreurs de regroupement du niveau *émission*.

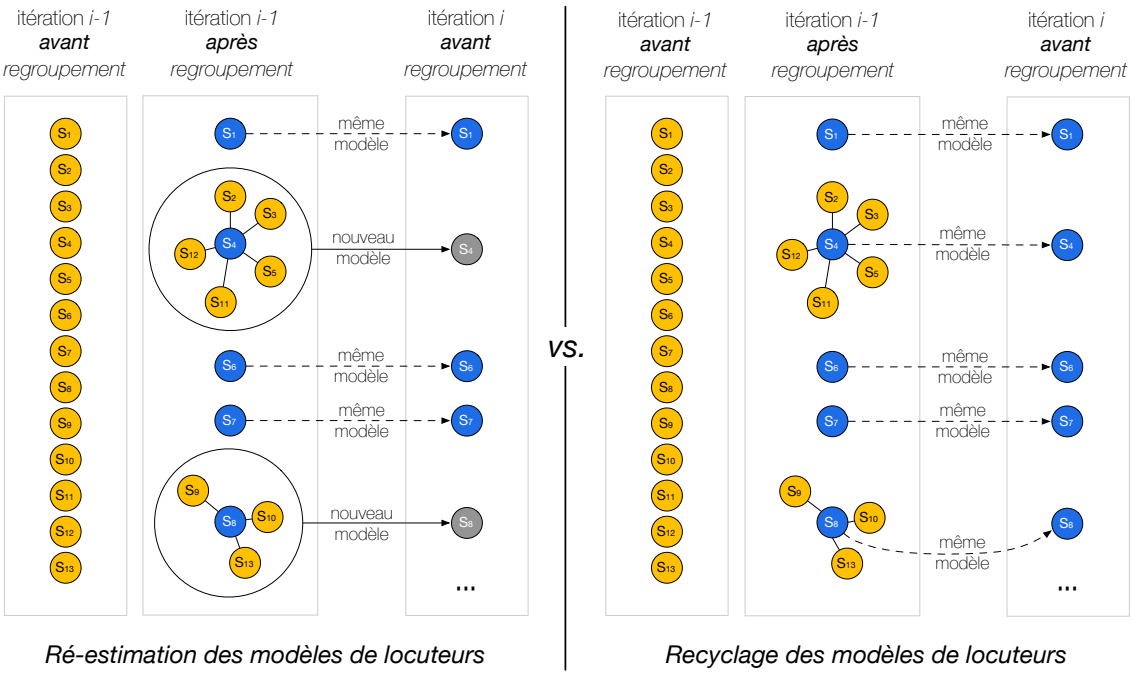

Figure *6.2 – Représentation schématique du procédé de* recyclage *des modèles de locuteurs entre deux itérations de l'architecture de regroupement incrémental.*

Nous proposons une solution (illustrée en figure 6.2) où, lors d'une itération i de notre architecture de regroupement incrémental, les modèles de locuteurs correspondant aux classes regroupées durant l'itération $i-1$ ne seraient pas recalculés. Nous n'avons expérimenté cette approche alternative qu'avec le regroupement ILP, pour lequel la notion de classe « centre » de regroupement est clairement définie. Cette classe centrale représente le modèle de locuteur le plus proche des autres en

termes de distance au sein d'un regroupement. Nous considérons donc que le modèle i-vector représentant la classe centrale d'un regroupement de l'itération $i-1$ est le plus représentatif de cette classe, et nous l'employons tel quel pour représenter cette classe regroupée lors de l'itération i (notion de « recyclage » des modèles de locuteurs). Cette approche permet non seulement de gagner du temps, car il n'est plus nécessaire d'apprendre de nouveaux modèles pour représenter les classes regroupées, mais aussi de limiter l'introduction de variabilité indésirable. Cette approche serait plus difficile à expérimenter avec le regroupement HAC, étant donné que cette notion de « centre » est absente. Il serait cependant possible de déterminer *a posteriori* le modèle i-vector le plus central d'une classe. Une autre possibilité serait, par exemple, d'utiliser le modèle i-vector correspondant à la classe à l'origine d'une succession de regroupements hiérarchique.

6.2.1 Expériences

Dans cette partie, nous présentons une confrontation des résultats obtenus entre les versions **avec** et **sans** implémentation du procédé de « recyclage ». Les expériences ont été menées sur les sept collections d'émissions du niveau *programme*, et l'approche de regroupement avec laquelle le procédé de *recyclage* a été expérimenté correspond à l'approche CC+ILP$_{\text{PLDA}}$. Nous présentons dans un premier temps les résultats en termes de DER$_{\text{de collections}}$ (tableau 6.3) et DER$_{\text{d'émissions}}$ (tableau 6.4) obtenus sur les sept collections étudiées. La confrontation de ces résultats à ceux obtenus avec l'approche « traditionnelle », où de nouveaux modèles de locuteurs sont calculés pour représenter les classes fusionnées, permet d'évaluer l'impact du procédé de *recyclage* sur la classification produite. Dans les tableaux 6.3 et 6.4, les DER présentés pour l'approche « traditionnelle » correspondent aux résultats déjà présentés dans les tableaux 6.1 et 6.2 de la partie précédente (en-têtes de colonnes de couleur verte dans les tableaux).

Nous présentons ensuite, en nous appuyant sur le tableau 6.5, une comparaison entre les durées relevées pour effectuer les deux versions du regroupement incrémental CC+ILP$_{\text{PLDA}}$ (avec et sans *recyclage*). Enfin, nous confrontons l'approche de regroupement incrémental à son alternative par regroupement global dans le (tableau 6.6).

▷ Comparaison en termes de DER

La différence entre les résultats moyens des deux versions du regroupement CC+ILP$_{\text{PLDA}}$ sur les sept collections d'émission du niveau *programme* n'est que très

légère. Ces résultats moyens ne sont cependant pas en faveur de l'implémentation du procédé de *recyclage* : le $DER_{\text{de collection}}$ moyen a augmenté de 0,95% (tableau 6.3), et le $DER_{\text{d'émissions}}$ moyen a augmenté de 0,24% (tableau 6.4). Le $DER_{\text{de collections}}$ moyen obtenu avec implémentation du procédé de *recyclage* (21,87%) reste cependant inférieur au $DER_{\text{de collections}}$ obtenu avec l'architecture de regroupement global, où l'approche $CC{+}ILP_{PLDA}$ permettait d'atteindre un score de 23,29%.

Configuration	**ILP_{PLDA} niveau émission**	**CC. + ILP_{PLDA} niveau collection**	**CC. + ILP_{PLDA} niveau collection + « Recyclage »**
Seuil β	-	-30	-30
Seuil δ	20	-30	-30
BFM Story	44,75%	14,32%	15,09%
Planète Showbiz	78,76%	47,35%	51,35%
Ça vous Regarde	35,73%	21,90%	22,08%
Entre les Lignes	85,77%	14,58%	15,34%
LCP Info	61,42%	20,86%	22,03%
Pile et Face	43,62%	23,41%	23,74%
Top Questions	57,20%	28,23%	28,88%
Moyenne	**54,83%**	**20,92%**	**21,87%**

Table *6.3* – **$DER_{\text{de collections}}$** *obtenus sur les collections de niveau* programme *avec l'architecture de regroupement incrémental pour les versions de l'approche de regroupement* $CC{+}ILP_{PLDA}$ **avec** *et* **sans** *« Recyclage ».*

Configuration	**ILP_{PLDA} niveau émission**	**CC. + ILP_{PLDA} niveau collection**	**CC. + ILP_{PLDA} niveau collection + « Recyclage »**
Seuil β	-	-30	-30
Seuil δ	20	-30	-30
BFM Story	10,70%	10,35%	10,77%
Planète Showbiz	37,41%	36,49%	37,10%
Ça vous Regarde	18,51%	18,51%	18,51%
Entre les Lignes	15,01%	14,24%	15,01%
LCP Info	9,60%	10,05%	9,38%
Pile et Face	20,05%	19,72%	20,05%
Top Questions	13,31%	14,24%	13,96%
Moyenne	**14,92%**	**14,79%**	**15,03%**

Table *6.4* – **$DER_{\text{d'émissions}}$** *obtenus sur les collections de niveau* programme *avec l'architecture de regroupement incrémental pour les versions de l'approche de regroupement* $CC{+}ILP_{PLDA}$ **avec** *et* **sans** *« Recyclage ».*

Le seul gain observé avec l'approche où les modèles de locuteurs sont *recyclés* est le $DER_{\text{d'émissions}}$ de la collection *LCP Info*, qui descend de 10,05% à 9,38%. L'intérêt de cette méthode semble donc discutable en termes de résultats, il apparaît plus

efficace de modéliser les classes issues d'un regroupement par un modèle de locuteur appris sur l'ensemble des données de ces classes.

▷ Comparaison en termes de durées

L'intérêt du procédé de *recyclage* réside essentiellement dans le gain obtenu en termes de durées de traitement. Le tableau 6.5 présente une comparaison des durées d'exécutions de l'architecture de regroupement incrémental, pour les sept collections d'émission du niveau *Programme*, et pour les deux versions du regroupement de niveau collection $CC{+}ILP_{PLDA}$ (l'une implémentant le procédé de *recyclage*, l'autre non). La dernière colonne de ce tableau présente les durées relevées pour effectuer le regroupement de niveau collection $CC{+}ILP_{PLDA}$ avec l'architecture de regroupement global. Toutes les durées présentées ont été mesurées dans les mêmes conditions en termes de matériel et de nombre de fils d'exécution (threads). Ces durées ne tiennent pas compte de l'étape de SRL d'émissions, où chaque enregistrement d'une collection est traité séparément et dont les segmentations produites servent d'entrée aux regroupements de niveau collection.

Architecture	**Regroupement incrémental**			**Rgpt. global**
Configuration	**CC. + ILP$_{PLDA}$**	**CC. + ILP$_{PLDA}$ + « Recyclage »**	**Taux de réduction**	**CC. + ILP$_{PLDA}$**
BFM Story	17h27m	2h43m	84,45%	31m26s
Planète Showbiz	27h57m	12h48m	54,21%	1h55m
Ça vous Regarde	3h57m	5m49s	97,55%	51s
Entre les Lignes	4h45m	3m25s	98,80%	41s
LCP Info	7h23m	36m17s	91,82%	11m9s
Pile et Face	2h55m	5m53s	96,64%	1m
Top Questions	1h40m	7m27s	92,56%	1m32s

Table *6.5 – Durée totale pour effectuer le regroupement incrémental, pour les versions de l'approche de regroupement $CC{+}ILP_{PLDA}$* **avec** *et* **sans** *« Recyclage », pour chaque collection étudiée.*

Il va de soi que la version du regroupement implémentant le procédé de *recyclage* est beaucoup moins coûteuse en temps, comparé à la version alternative où les modèles de locuteurs correspondant aux classes regroupées lors d'une itération $i-1$ sont recalculés pour effectuer l'itération i. Le gain de temps observé à ce propos est considérable, comme en témoigne la troisième colonne du tableau 6.5, où sont présentés les taux de réduction entre les durées d'exécution des deux versions du regroupement. Ce taux de réduction est fonction non seulement du volume de la collection, qui conditionne directement le nombre d'itérations de l'approche de regroupement incrémental, mais également de la durée des émissions de la collection, et indirectement du nombre de classes de locuteurs, qui influent sur la complexité de

l'étape de regroupement $CC{+}ILP_{PLDA}$. Le procédé de *recyclage* des modèles de locuteurs permet donc de réduire la durée nécessaire pour l'exécution du regroupement incrémental de manière considérable, tout en permettant de conserver des résultats en termes de DER satisfaisants, voire, quelque peu meilleurs.

Toutefois, la durée requise pour traiter intégralement les collections par regroupement incrémental reste nettement supérieure à celle du traitement par regroupement global. La dernière colonne du tableau 6.5 permet de constater l'ampleur des écarts de durées d'exécution entre les deux architectures de regroupement : l'approche de regroupement global est plus rapide que l'approche de regroupement incrémental, en particulier pour les collections volumineuses. À titre d'exemple, sur la collection *Planète Showbiz*, l'approche de regroupement incrémental avec implémentation du procédé de *recyclage* a duré environ 13 heures, contre seulement 2 heures pour l'approche de regroupement global.

L'intérêt de notre architecture de regroupement incrémental réside essentiellement dans le traitement des collections dynamiques, dont le volume augmenterait au cours du temps. Nous présentons en figure 6.3 une représentation des durées mesurées pour chaque itération du procédé de regroupement incrémental **avec** Recyclage, pour chacune des collections étudiées (l'échelle des ordonnées est logarithmique).

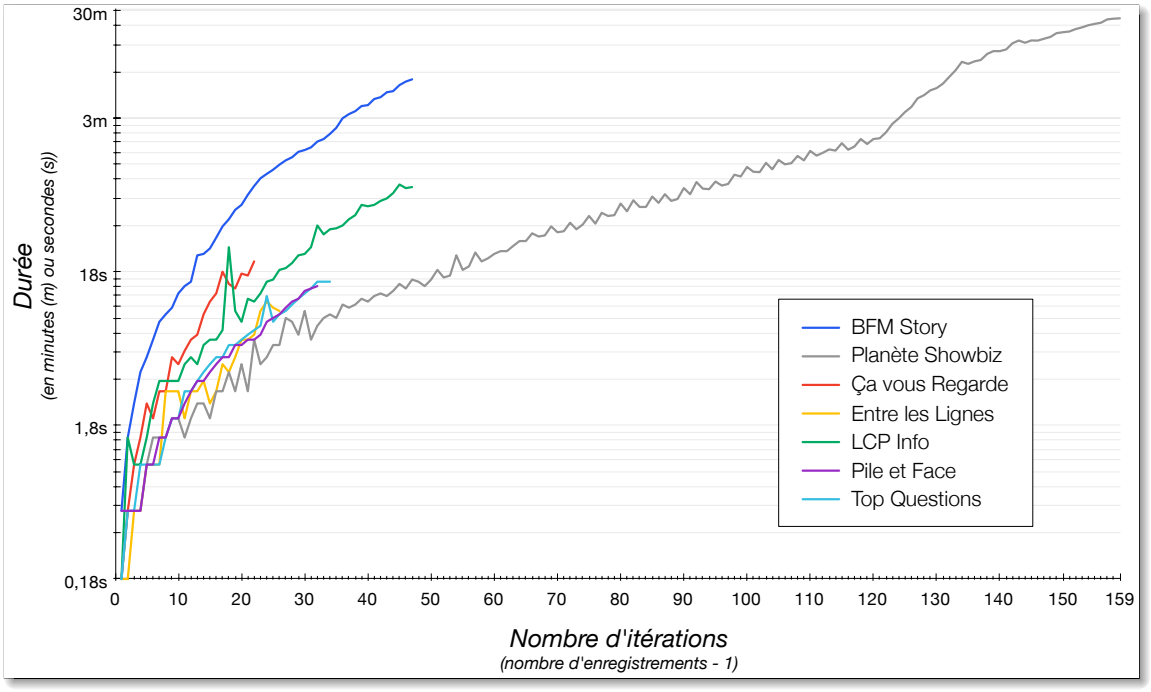

Figure *6.3 – Évolution des durées de traitement en fonction du nombre d'itérations pour chacune des collections étudiées (échelle logarithmique sur l'axe des ordonnées).*

Attention, la figure 6.3 présente bien les durées relevées pour chacune des itérations. La durée totale du procédé de regroupement incrémental, pour chacune des

collections étudiées (qui a déjà été présentée dans le tableau 6.5), correspondrait à la somme des durées relevées pour chacune des itérations.

On constate que si les premières itérations sont rapidement effectuées, les suivantes deviennent de plus en plus longues. Ce constat est néanmoins à nuancer dans la mesure où l'étape la plus coûteuse en temps concerne l'estimation des scores PLDA entre les différentes paires de classes. Cette observation, déjà évoquée dans la partie 5.4, a des conséquences néfastes sur l'architecture de regroupement global, qui agglomère et classifie itérativement les segmentations de chacun des enregistrements d'une collection. Étant donné que le nombre de classes impliquées dans les regroupements augmente au fil des itérations, la durée nécessaire pour estimer les scores PLDA entre les modèles de locuteurs correspondants augmente également. Or, l'estimation des scores PLDA est une étape indispensable effectuée à chaque itération, il n'est donc pas surprenant que l'approche de regroupement incrémental soit moins performante en termes de durée d'exécution totale, comparée à l'approche de regroupement global. En revanche, si l'on ne considère que la dernière itération des regroupements de types incrémentaux, on constate un gain important en termes de durées d'exécution (*cf.* tableau 6.6) :

Architecture	**Regroupement incrémental : itération finale**			**Rgpt. global**	
Configuration	**n^{bre} classes**	**CC. + ILP_{PLDA}**	**CC. + ILP_{PLDA} + « Recyclage »**	**n^{bre} classes**	**CC. + ILP_{PLDA}**
BFM Story	1909	47m29s	10m51s	2845	31m26s
Planète Showbiz	2552	43m47s	26m39s	4224	1h55m
Ça vous Regarde	645	19m34s	41s	814	51s
Entre les Lignes	435	22m4s	21s	732	41s
LCP Info	1065	20m19s	2m9s	1812	11m9s
Pile et Face	539	10m40s	29s	868	1m
Top Questions	545	4m55s	37s	1000	1m32s

Table *6.6 – Durées pour effectuer la dernière itération du regroupement incrémental, pour les versions de l'approche de regroupement CC+ILP_{PLDA}* **avec** *et* **sans** *« Recyclage », pour chaque collection étudiée.*

Comparer la durée de la dernière itération du regroupement incrémental à la durée du regroupement global revient à se placer dans un contexte applicatif où un nouvel enregistrement aurait été ajouté à la collection. Avec le regroupement global, il aurait été nécessaire d'effectuer un nouveau regroupement global sur l'ensemble des classes des segmentations produites au niveau *émission* de l'architecture. Avec le regroupement incrémental, on effectuerait une nouvelle itération où le nombre de classes impliquées serait moindre, du fait des regroupements effectués lors des itérations précédentes. La durée requise pour estimer les scores PLDA, lors de cette nouvelle itération, serait alors inférieure à celle d'un regroupement global.

6.2.2 Discussion

Le procédé de *recyclage* permet de viabiliser notre approche de regroupement incrémental en permettant de réduire la complexité de l'étape d'estimation des scores PLDA, qui correspond à l'étape la plus longue de nos approches pour le traitement de collections. Le traitement est plus long du fait de la multiplication des regroupements, en revanche, le nombre de classes impliquées dans les problèmes de classification est moins important du fait des regroupements itératifs. Cette approche permettrait donc de repousser le problème rencontré avec l'architecture de regroupement global quant à la complexité des problèmes ILP soumis à l'outil de résolution, où aucune solution ne semble pouvoir être déterminée en un laps de temps raisonnable si la combinaison de seuils β et δ mène à des problèmes impliquant trop de classes.

Le procédé de *recyclage* des modèles de locuteurs pourrait également être employé au niveau de la concaténation des segmentations d'émissions. Actuellement, avec cette architecture de regroupement incrémental, mais aussi avec l'architecture de regroupement global présentée dans le chapitre précédent, nous extrayons de nouveaux modèles de locuteurs pour représenter les classes regroupées lors de la dernière étape de regroupement du système de SRL d'émissions. Or la paramétrisation employée est identique : le dernier regroupement effectué au niveau *émission* est un regroupement ILP où les classes sont représentées par des modèles i-vectors de dimension 300. Nos approches de regroupement *globales*, ILP, HAC et leurs combinaisons avec l'approche de décomposition en composantes connexes (CC), manipulent donc exactement les mêmes types de modèles de locuteurs. Nous n'avons toutefois pas expérimenté cette généralisation, car à l'origine, la modélisation employée était différente entre les niveaux de regroupements *émission* (modèles GMM) et *collection* (modèles i-vector) Dupuy et al. [2012a].

6.3. Amorçage du procédé incrémental

Nous présentons dans cette partie une étude menée suite à nos observations sur le procédé de *recyclage* des modèles de locuteurs pour l'architecture de regroupement incrémental. L'architecture de regroupement global semble plus performante pour traiter une collection d'enregistrement, en revanche, si des enregistrements venaient enrichir cette même collection, il pourrait alors être judicieux de « continuer » le traitement avec l'architecture de regroupement incrémental. En effet, considérons qu'aujourd'hui nous disposons d'un ensemble d'enregistrements pouvant être assimilé à une collection. Aujourd'hui, à moins que cette collection ne soit trop volumineuse,

nous pourrions la traiter par l'intermédiaire de l'architecture de regroupement global, telle que présentée dans le chapitre 5. En revanche, si demain et dans les jours qui suivent, de nouveaux enregistrements devaient être amenés à enrichir la collection traitée, plusieurs stratégies s'offriraient à nous, en particulier, une combinaison entre les architectures de regroupement global et incrémental pourraient être envisagées. En fonction du contexte applicatif et des moyens à disposition, ces stratégies pourraient être (entre autres) :

- Si la fréquence d'acquisition entre les nouveaux enregistrements est faible et constante, nous pourrions avoir recours à l'architecture de regroupement global en considérant la totalité des enregistrements qui composent la collection, ceux ayant déjà fait l'objet d'un traitement ainsi que ceux ayant été nouvellement acquis.
- Dans le cas contraire, si la durée entre l'acquisition de deux enregistrements n'est pas suffisamment importante pour effectuer un regroupement global, il pourrait être acceptable d'attendre d'avoir accumulé un *ensemble* de nouveaux enregistrements et, de temps à autre, mettre à jour à la collection par regroupement global en incluant cet ensemble aux enregistrements déjà traités.
- Toujours dans un contexte où la durée entre l'acquisition de deux enregistrements ne serait pas suffisante pour effectuer un regroupement global, il pourrait être important de disposer rapidement des regroupements correspondant aux nouveaux enregistrements, auquel cas il ne serait pas envisageable d'accumuler les nouveaux enregistrements avant de les inclure à la collection déjà traitée. Dans ce cas, nous pourrions poursuivre le traitement avec l'architecture de regroupement incrémental, en considérant la segmentation de collection produite par l'architecture de regroupement global, avant acquisition de nouveaux enregistrements, comme point d'entrée pour le regroupement incrémental.

Dans cette partie, nous considérons le dernier point évoqué et expérimentons une *variante* de l'architecture de regroupement incrémental où une collection dite *initiale* serait d'abord traitée par regroupement global. Cette variante, illustrée en figure 6.4, permet donc de simuler un contexte applicatif concret.

Nous souhaitons observer le comportement des approches de classification, au travers des résultats en termes de DER et en nombre de locuteurs récurrents correctement détectés, et effectuer une comparaison avec les résultats obtenus avec un regroupement global sur l'ensemble des données expérimentées.

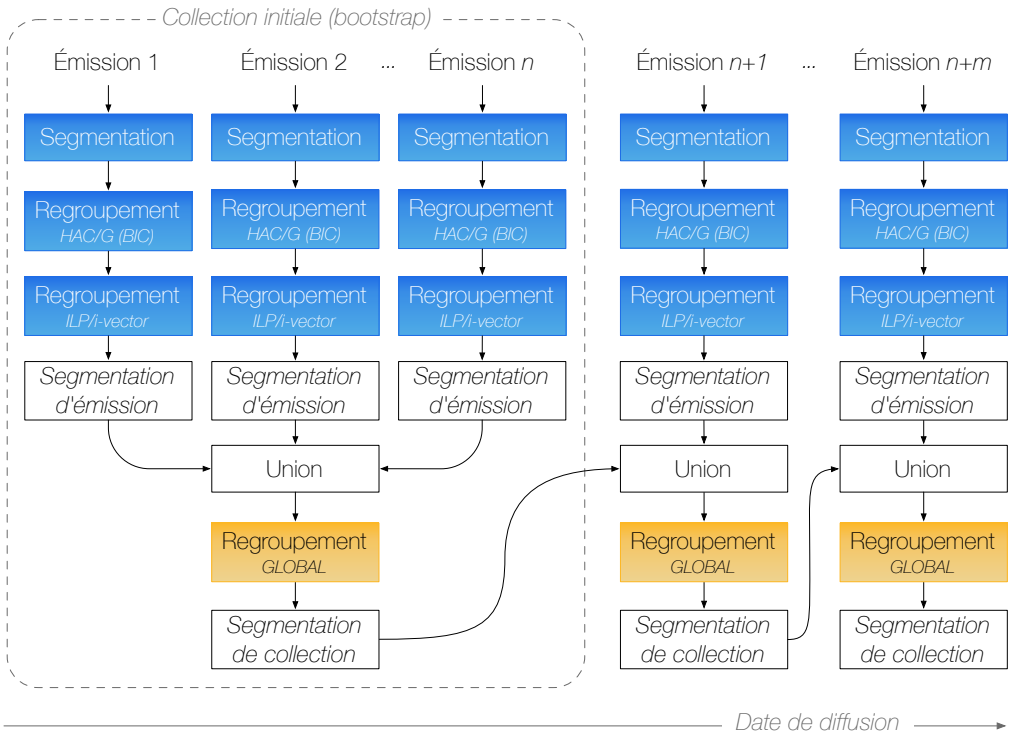

Figure *6.4 – Architecture par regroupement incrémental avec collection* bootstrap *pour la SRL de collections.*

6.3.1 Expériences

Pour mener à bien cette étude, nous n'avons considéré que la plus volumineuse de nos collections, c'est-à-dire, celle regroupant l'ensemble de nos données d'évaluation (la collection d'émissions de niveau *thématique*). Nous avons expérimenté l'architecture présentée en figure 6.4 à partir de collections *initiales* de différentes tailles, afin de mieux constater leur impact sur les résultats obtenus. Ces collections initiales, dont les caractéristiques sont présentées dans le tableau 6.7, sont constituées des n premiers enregistrements, selon l'ordre chronologique de diffusion, de la collection d'émissions de niveau *Thématique*. Ces collections initiales sont traitées par regroupement global, et les enregistrements restants sont itérativement traités (ajoutés) par regroupement incrémental.

Étant donné les observations faites à propos des durées de traitement de l'approche de regroupement incrémental, nous n'employons que le regroupement $CC{+}ILP_{PLDA}$ implémentant le procédé de *recyclage* des modèles de locuteurs. Le regroupement global est, par conséquent, effectué avec l'approche de regroupement $CC{+}ILP_{PLDA}$ et les seuils β, pour la décomposition en composantes connexes, et δ, pour la résolution du problème ILP_{PLDA}, sont toujours fixés à -20 (il s'agit du seuil optimal sur la collection d'émissions *Thématique* (qui regroupe l'ensemble des

Regroupement global				Regroupement incrémental		
Collection initiale	n^{bre} enr.	Durée		n^{bre} enr. restants	Durée	
		audio	UEM		audio	UEM
Boot. 0%	0	0h0m	0h0m	374	177h48m	67h13m
Boot. 25%	97	44h28m	18h55	277	133h19m	48h19m
Boot. 50%	216	89h52m	39h18m	158	88h46m	27h56m
Boot. 75%	303	133h38m	57h33m	71	44h10m	9h41m
Boot. 100%	374	177h48m	67h13m	0	0h0m	0h0m

Table *6.7 – Constitution des collections initiales.*

données) avec l'approche de regroupement global $CC{+}ILP_{PLDA}$).

Malgré les mesures mises en place pour minimiser la durée du regroupement incrémental, les expériences menées avec la collection d'émissions thématique sont très longues, à tel point que nous sommes contraints de ne présenter que des résultats partiels. La principale source de ce problème de durées est due à l'implémentation de l'outil permettant d'estimer les scores PLDA, qui n'est pas optimisé pour ce genre de traitements. Comme expliqué dans la partie précédente, qui traite du procédé de *recyclage* des modèles de locuteurs, il ne s'agit que d'un problème d'implémentation, l'étape d'estimation des scores PLDA en soi n'est pas censée être coûteuse. Nous ne nous étions d'ailleurs pas heurtés à ce problème lors de nos premiers travaux sur le regroupement incrémental à partir de collections initiales. Ces travaux, présentés en annexe D, avaient été effectués sur une collection à peine moins volumineuse (310 enregistrements) et pourtant beaucoup plus rapidement (environ 6 heures). La configuration n'était cependant pas la même : nous utilisions le regroupement ILP avec des modèles i-vector de dimension 50, et la mesure de similarité entre les modèles était estimée avec la distance de *Mahanalobis* sans avoir recours à l'outil *IvTest* de la suite *Alize*.

Les résultats obtenus pour les itérations effectuées sont présentés en figures 6.5 pour le $DER_{\text{de collections}}$ et 6.6 pour le $DER_{\text{d'émissions}}$. Les résultats obtenus à partir des différentes collections initiales ont été superposés afin de mieux observer le comportement de l'approche de regroupement incrémental vis-à-vis des données initialement traitées par regroupement global (les DER obtenus par regroupement global pour les collections initiales sont symbolisés par des droites sur les deux figures).

Concernant le traitement des collections initiales par regroupement global, les constatations effectuées dans le chapitre précédent quant à la difficulté de la SRL de collections semblent se confirmer. Plus la collection est fournie en enregistrements, plus les DER sont élevés : les $DER_{\text{de collections}}$ varient entre 19,61% pour *Boot.25%* et 26,24% pour *Boot.100%*, et les $DER_{\text{d'émissions}}$ varient entre 13,11% *(Boot.25%)*

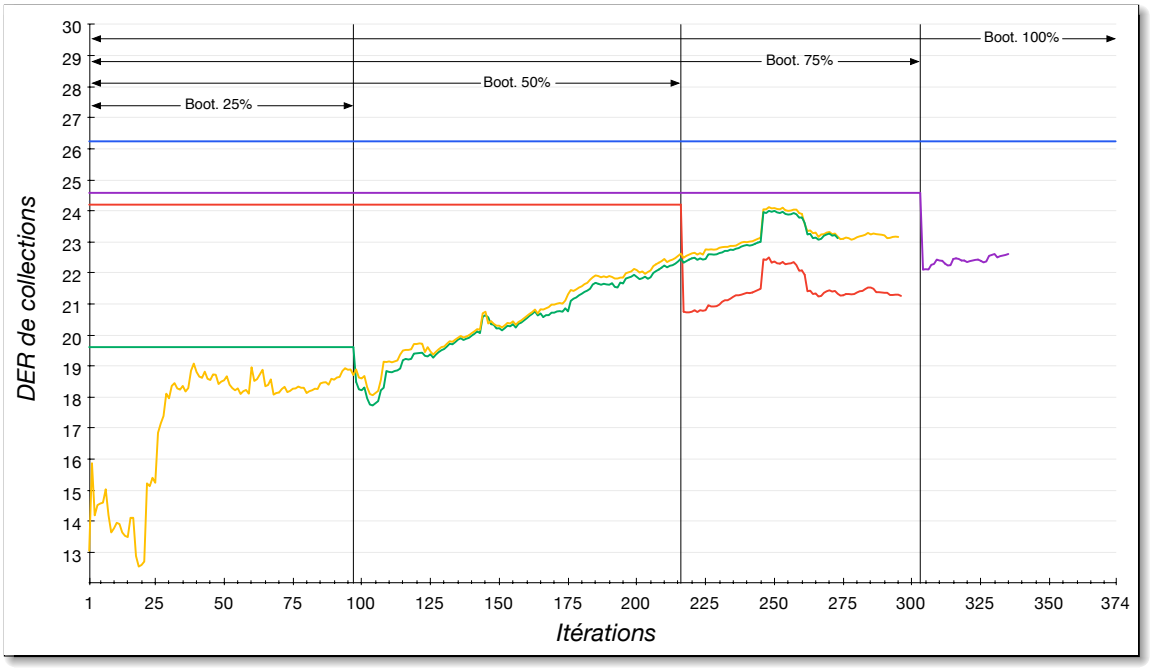

Figure *6.5 – Évolution des* **$\mathbf{DER}_{\text{de collections}}$** *en fonction de la collection initiale sur l'ensemble des enregistrements.*

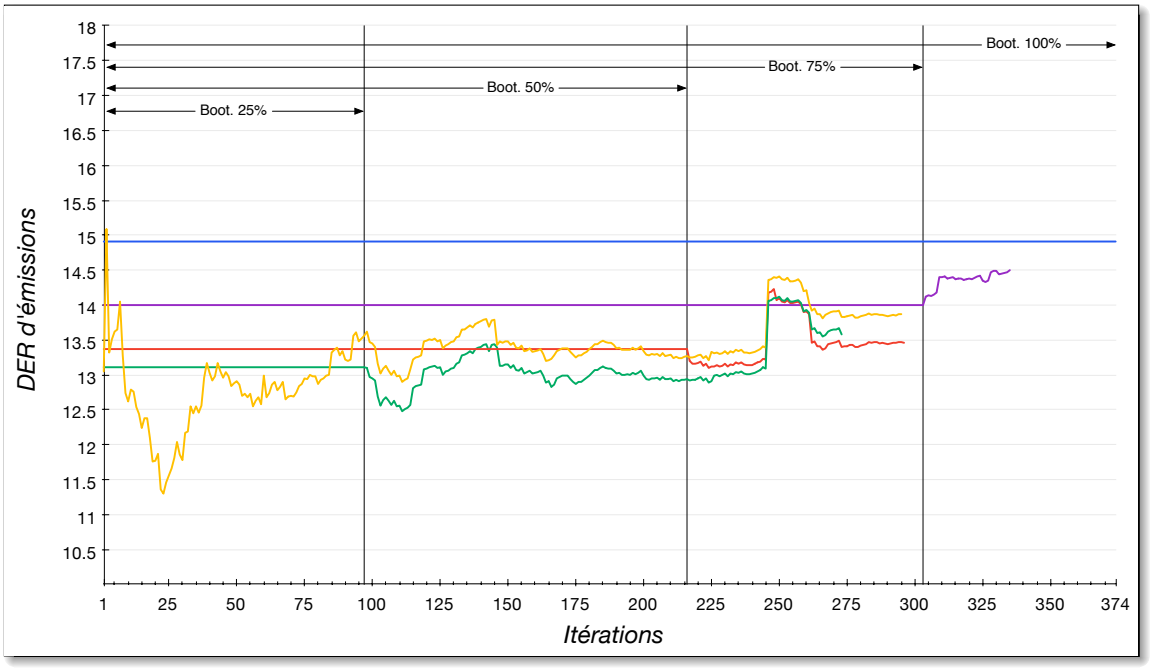

Figure *6.6 – Évolution des* **$\mathbf{DER}_{\text{d'émissions}}$** *en fonction de la collection initiale sur l'ensemble des enregistrements.*

et 14,91% *(Boot.100%)*.

Les taux d'erreur mesurés lors des différentes itérations ont globalement tendance à augmenter, même si nous observons des variations plus ou moins importantes au fil des itérations. La superposition des résultats obtenus avec les différentes collections initiales permet de constater que ces dernières n'affectent pas considérablement les

regroupements effectués entre les itérations : à partir de l'itération n°97, les courbes représentant les $DER_{\text{de collections}}$ obtenus à partir des collections initiales *Boot.0%* (collection initiale vide, le procédé de regroupement incrémental démarre avec le premier enregistrement) et *Boot.25%* se superposent presque parfaitement. La même observation peut être effectuée entre les collections initiales *Boot.0%* et *Boot.50%* (à partir de l'itération n°216), où malgré un décalage d'environ 2%, les courbes représentant les $DER_{\text{de collections}}$ se ressemblent beaucoup. Ces remarques valent également pour les $DER_{\text{d'émissions}}$.

6.3.2 Discussion

Nous pouvons constater, à l'aide de la figure 6.5, un comportement inattendu au niveau des $DER_{\text{de collections}}$ obtenus lors des premières itérations suivant le regroupement global d'une collection initiale. On constate une diminution importante des $DER_{\text{de collections}}$, dont la plus flagrante a lieu avec la collection initiale *Boot.50%* : la diminution observée est de 3,45%, le $DER_{\text{de collections}}$ chute de 24,2% à 20,75%. Les premières analyses effectuées à ce propos suggèrent que cette diminution du $DER_{\text{de collections}}$ n'est pas due à l'ajout d'un nouvel enregistrement à la collection initiale, mais plutôt au nouveau regroupement effectué sur les données. Nous avons pu observer une diminution du $DER_{\text{de collections}}$ similaire en effectuant, à nouveau, un regroupement global sur les enregistrements de la collection initiale *Boot.50%*. Plus simplement, deux regroupements globaux ont été enchaînés sur les mêmes données :

- Le premier regroupement global est identique à celui présenté dans le chapitre précédent. Les segmentations d'émissions des enregistrements composant la collection ont été réunies au sein d'un unique fichier, et des modèles de locuteurs i-vector ont été calculés sur les données de chaque classe pour effectuer le regroupement global. Le $DER_{\text{de collections}}$ est égal à 24,2%.
- Le deuxième regroupement global démarre avec la segmentation de collection produite à l'issue du premier regroupement global (aucun nouvel enregistrement n'est ajouté). Le $DER_{\text{de collections}}$ obtenu à l'issue de ce deuxième regroupement global est de 20,86%, donc inférieur de 3,34% (ce qui correspond approximativement à la diminution observée lors de l'ajout du premier enregistrement à la collection *Boot.50%*).

La diminution observée n'est donc pas imputable à l'ajout d'un nouvel enregistrement, ni à l'implémentation du procédé de *recyclage* des modèles de locuteurs. En effet, que le procédé de *recyclage* soit implémenté ou non, une diminution significative du $DER_{\text{de collection}}$ est observée. Nos premières analyses suggèrent que la raison des

diminutions observées est due à l'enchaînement des décompositions en composantes connexes. Nous illustrons l'effet produit avec l'exemple présenté en figure 6.7.

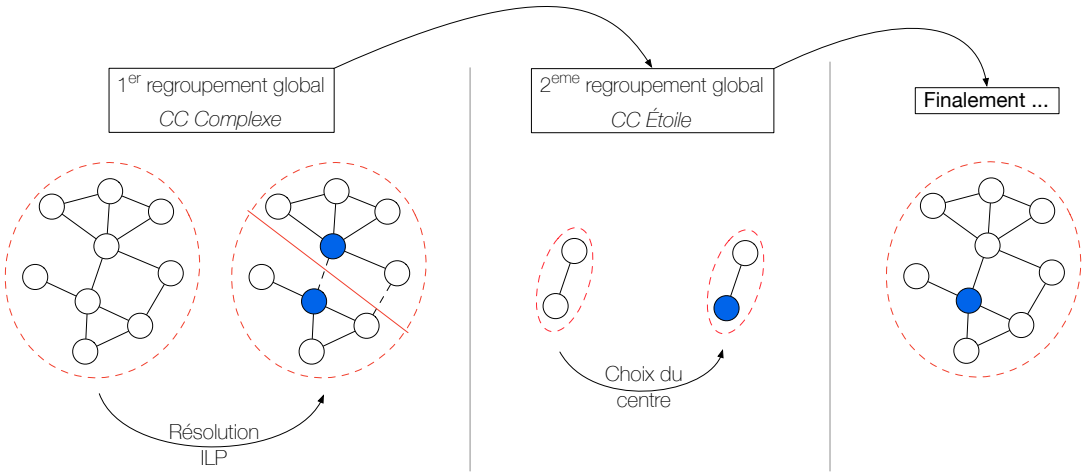

Figure *6.7 – Illustration de l'effet produit par l'enchaînement de deux décompositions en composantes connexes. Les sommets représentent les classes, les arêtes sont correspondent à des scores PLDA inférieurs au seuil δ. Les sommets bleu représentent les centres.*

Dans notre exemple, l'approche de décomposition en composantes connexes produit une composante connexe *complexes* lors du premier regroupement global. Celle-ci est résolue par l'approche de regroupement ILP, qui découpe la CC complexe en deux classes. Lors du regroupement global supplémentaire, ces deux classes peuvent être regroupées, leur score de vraisemblance étant inférieur au seuil δ. Finalement, toutes les classes de la composante connexe complexe du premier regroupement global auront été regroupées, ce qui pose problème dans la mesure où le score PLDA entre certaines des classes était supérieur au seuil δ. La contrainte ILP portant sur la *distance* entre le centre d'une classe et les autres éléments de la classe n'est donc plus respectée. S'agit-il d'un effet indésirable ou d'une piste vers un apport bénéfique ? Quoi qu'il en soit, nous déduisons de cet *effet* qu'il n'existe pas de seuil δ optimal pour l'ensemble de l'espace.

6.4. Analyse et bilan

Nous présentons, dans cette partie, une analyse relativement synthétique sur l'approche de regroupement incrémental pour le traitement des collections de documents audiovisuels. Les observations et constatations formulées reposent sur une analyse intermédiaire et complémentaire de l'application de ce procédé incrémental sur les différentes collections proposées à l'étude. Cette analyse intermédiaire est disponible en annexe B.

L'approche de regroupement incrémental expérimentée dans le cadre de cette thèse repose sur les mêmes techniques que celles mises en œuvre pour l'approche de regroupement global : d'une part l'approche permettant la décomposition d'un problème de classification volumineux en plusieurs sous-problèmes indépendants, dont la plupart sont triviaux à résoudre; d'autre part, les approches de regroupement ILP et HAC, utilisées pour résoudre les sous-problèmes plus *complexes*. Paramétrisation, modélisation en locuteurs et méthode de scoring sont également les mêmes que précédemment.

Nous avons également introduit une approche visant à alléger le procédé de regroupement incrémental en minimisant le nombre de modèles i-vectors à extraire lors de chaque itération (*recyclage* des modèles de locuteurs) : seuls les modèles correspondant aux classes de locuteurs de l'enregistrement ajouté doivent être extraits, les autres sont récupérés depuis les itérations précédentes. Pour les classes de locuteurs correspondant à des regroupements, lors des itérations précédentes, le modèle i-vector sélectionné pour représenter la classe regroupée correspond à celui le plus *central* du regroupement.

6.4.1 Méthode de classification

Les résultats obtenus sur les collections d'émissions de niveau *Programme* et sur les collections *Temporelles*, par les méthodes de classification $CC{+}ILP_{PLDA}$ (sans recourir au procédé de *recyclage*) et $CC{+}HAC_{PLDA}$ sont extrêmement proches, en particulier pour ce qui est des $\text{DER}_{\text{de collections}}$ et des proportions de locuteurs récurrents correctement détectées.

La méthode de classification $CC{+}ILP_{PLDA}$ permet d'obtenir un $\text{DER}_{\text{de collections}}$ moyen légèrement inférieur, sur les collections d'émissions du niveau *Programme*, à celui de la méthode $CC{+}HAC_{PLDA}$ (20,92% contre 21,38%). Les deux méthodes de classification donnent cependant des $\text{DER}_{\text{de collections}}$ absolument identiques pour toutes les collections du niveau *Programme*, excepté pour *BFM Story*. Il n'est pas surprenant de constater une telle différence entre les $\text{DER}_{\text{de collections}}$ moyens étant donné que les enregistrements de la collection *BFM Story* représentent à eux seul un tiers de la totalité des données évaluées (cette collection pèse donc énormément dans le calcul de la moyenne des $\text{DER}_{\text{de collections}}$, qui est pondérée par la durée des collections). Concernant les collections *Temporelles*, la similitude entre les $\text{DER}_{\text{de collections}}$ obtenus par les méthodes de classification est encore plus flagrante : les $\text{DER}_{\text{de collections}}$ obtenus, en moyenne, sont de 20,69% pour la méthode $CC{+}ILP_{PLDA}$, et 20,71% pour la méthode $CC{+}HAC_{PLDA}$. La seule collection pour laquelle les $\text{DER}_{\text{de collections}}$ ne sont

pas identiques entre les deux méthodes de classification est la collection *Temporelle* n°8.

Les $DER_{d'émissions}$ fluctuent légèrement selon la méthode de classification employée. La différence dépend des collections étudiées. Avec les collections d'émissions de niveau *Programme*, cette différence est, au plus, de 0,92%. Avec les collections *Temporelles*, la plus grande différence observée n'est que de 0,09% (les $DER_{d'émissions}$ moyens sont cependant identiques avec les deux méthodes de classification).

En ce qui concerne la proportion de locuteurs récurrents correctement détectés par les deux méthodes de classification, les résultats sont également très proches, comme en témoigne les figures 6.8 (collections d'émissions de niveau *Programme*) et 6.9 (collections *Temporelles*). Si l'on peut observer d'infimes variations sur les collections d'émissions, on constate que les deux méthodes de classification permettent de détecter exactement les mêmes proportions de locuteurs récurrents sur les collections *Temporelles* (*cf.* annexeB).

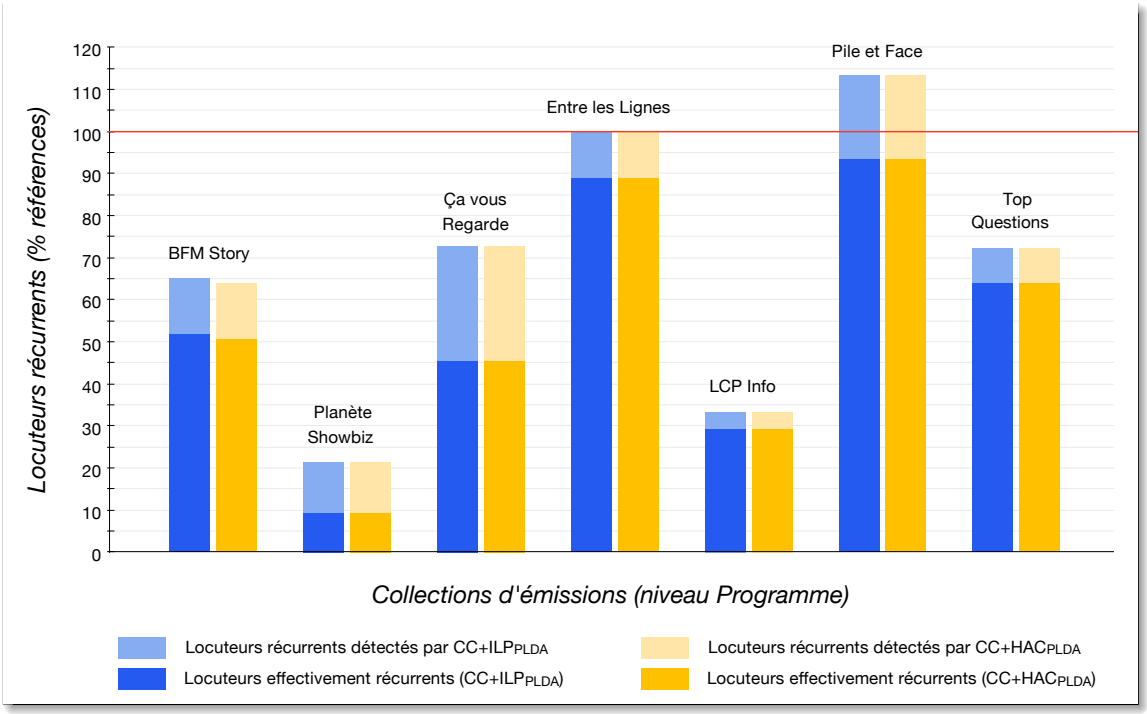

Figure *6.8 – Représentation du nombre de locuteurs récurrents détectés par les approches de regroupement* $CC+ILP_{PLDA}$ *et* $CC+HAC_{PLDA}$*, pour chacune des collections d'émissions étudiées. La partie foncée des barres représente la proportion de locuteurs récurrents détectés par les approches de regroupement qui sont effectivement récurrents d'après les segmentations de référence.*

▷ Comparaison avec l'approche de regroupement global

Concernant les collections d'émissions de niveau *Programme*, les $DER_{de\ collections}$ obtenus avec l'approche incrémentale sont inférieurs à ceux obtenus avec l'approche

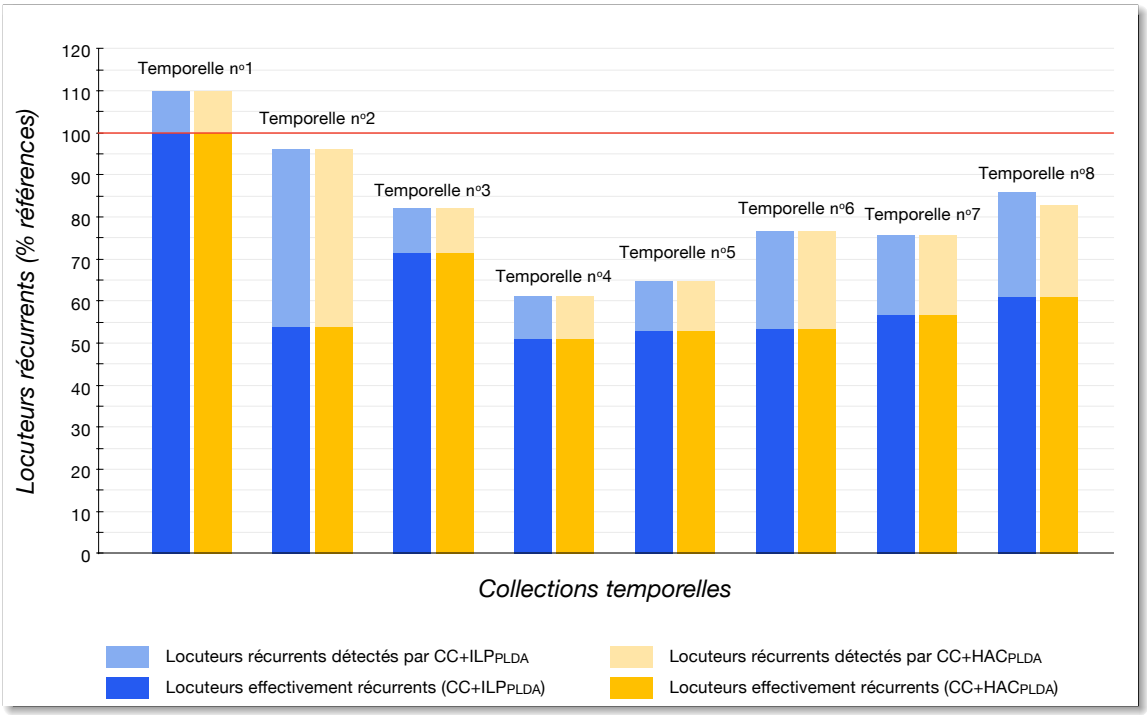

Figure *6.9 – Représentation du nombre de locuteurs récurrents détectés par les approches de regroupement* CC+ILP$_{PLDA}$ *et* CC+HAC$_{PLDA}$*, pour chacune des collections temporelles étudiées. La partie foncée des barres représente la proportion de locuteurs récurrents détectés par les approches de regroupement qui sont effectivement récurrents d'après les segmentations de référence.*

de regroupement global : la méthode de classification *CC+ILP$_{PLDA}$* permet d'atteindre, en moyenne, en incrémental, 20,92% (contre 23,29% pour l'approche de regroupement global); la méthode *CC+HAC$_{PLDA}$* permet d'atteindre, en moyenne, en incrémental, 21,38% (contre 23,18% pour l'approche de regroupement global). Bien que les deux méthodes de classification donnent des résultats semblables avec l'approche de regroupement incrémental, la méthode *CC+ILP$_{PLDA}$* semble préférable, contrairement à l'approche de regroupement global où la méthode *CC+HAC$_{PLDA}$* s'est révélée être plus efficace. Avec les collections *Temporelles*, les DER$_{\text{de collections}}$ moyens obtenus avec les approches de regroupement incrémental et global sont quasiment identiques pour la méthode de classification *CC+ILP$_{PLDA}$* (la différence n'est que de 0,03%).

Parmi les locuteurs récurrents détectés par les méthodes de classification, c'est-à-dire, les classes de locuteurs constituées de segments provenant d'au moins deux enregistrements différents, la proportion qui correspond effectivement à des locuteurs récurrents d'après les segmentations de référence est légèrement plus élevée avec l'approche de regroupement incrémental (par comparaison avec l'approche de regroupement global). En revanche, le nombre total de locuteurs récurrents correctement détecté est, en moyenne, plus élevé avec l'approche de regroupement global. L'approche de regroupement incrémental permet donc de faire moins d'erreurs que

l'approche de regroupement global, sur la détection des locuteurs récurrents, mais ne permet pas d'en détecter autant.

6.4.2 Recyclage et collection initiale

Nous n'avons expérimenté le procédé de *recyclage* qu'avec la méthode de classification $CC{+}ILP_{PLDA}$, qui présente la particularité de regrouper les classes autour d'un « centre » (une classe centrale). Ce procédé permet d'économiser du temps lors de chaque itération du regroupement incrémental, cependant, les résultats en termes de $\text{DER}_{\text{de collections}}$ et en proportion de locuteurs récurrents correctement détectés se dégradent légèrement, comparés aux résultats obtenus avec la même approche n'implémentant pas le procédé de *recyclage*.

Ces détériorations sont toutefois minimes au regard du gain de temps obtenu, en particulier dans le contexte particulier où des enregistrements sont ajoutés à une collection initiale. Nous avons pu observer, dans ce cas précis, un comportement pour le moins intéressant : la première itération du regroupement incrémental, faisant suite au regroupement global d'une collection initiale, permettrait de diminuer le $\text{DER}_{\text{de collections}}$ obtenu sur la collection initiale par regroupement global (la plus flagrante diminution observée étant d'environ 3%). Les premières investigations à ce propos suggèrent que cette diminution est due au procédé de *recyclage*, des expériences complémentaires sont cependant nécessaires pour valider cette théorie.

6.4.3 Discussion générale

Le regroupement incrémental que nous avons proposé dans le cadre de ce manuscrit, malgré le problème rencontré quant aux durées excessives de l'outil *IvTest* pour estimer les scores PLDA, a permis d'atteindre des résultats satisfaisants en termes de $\text{DER}_{\text{de collections}}$. Ces travaux sur l'approche de regroupement incrémental peuvent être considérés comme une extension de l'étude préliminaire présentée en annexe D, et sont, à ce titre, toujours en cours d'investigation.

Troisième partie

Conclusions et perspectives

CHAPITRE 7

Conclusions et perspectives

Les travaux présentés dans cette thèse s'inscrivent dans le cadre de la tâche de segmentation et regroupement en locuteurs (SRL) pour le traitement de collections de documents audiovisuels. Une collection est constituée d'un ensemble d'enregistrements audiovisuels où interviennent plusieurs locuteurs dont certains, les locuteurs *récurrents*, présentent la particularité d'intervenir dans plusieurs enregistrements de la collection. Les travaux menés dans cette thèse visent à répondre à deux questions : comment adapter la SRL pour le traitement des collections, et ainsi être en mesure de détecter les locuteurs récurrents ? Comment rendre les approches proposées robustes face à de gros volumes de données, ou aux collections susceptibles d'évoluer au cours du temps ?

La tâche de SRL consiste à déterminer automatiquement les interventions des différents locuteurs dans un enregistrement audiovisuel. Nos travaux quant à l'adaptation de cette tâche pour le traitement des collections ont principalement porté sur les approches de regroupement – ou classification – en locuteurs, de manière à déterminer, en particulier, les interventions des locuteurs récurrents.

L'autre aspect des travaux présentés dans cette thèse concerne le traitement des collections *volumineuses*. La quantité de données audiovisuelles disponible ne cesse de croître : en juin 2014 YouTube annonçait la mise en ligne de 100 heures de vidéo chaque minute. En février 2015 ce sont désormais 300 heures de vidéo qui sont publiées chaque minute. Or, les approches traditionnellement employées pour la classification en locuteurs ne sont pas adaptées au traitement de gros volumes de données. Dans les travaux que nous avons présentés, nous avons tenté d'apporter des solutions en proposant des architectures et des méthodes de regroupement robustes face aux grandes quantités de données.

7.1. Conclusion

La principale différence entre SRL d'émissions et SRL de collections repose sur la détection des locuteurs récurrents. La SRL d'émissions en est incapable dans la mesure où les enregistrements sont traités indépendamment les uns des autres, elle permet cependant d'obtenir de très bons résultats de classification sur les enregistrements. Afin de détecter les locuteurs récurrents, il est nécessaire de considérer l'ensemble des enregistrements d'une collection simultanément. Cette étape peut se faire à différents niveaux de l'architecture, cependant, plus elle est réalisée tôt dans la chaîne de traitement, plus le temps de calcul est important. L'approche communément admise pour traiter une collection consiste à traiter, dans un premier temps, les différents enregistrements de la collection par un système de SRL d'émissions, puis effectuer une étape de regroupement supplémentaire en considérant l'ensemble des segmentations d'émissions ainsi obtenues.

Cette étape de regroupement supplémentaire peut être réalisée de deux manières différentes : globalement, ou incrémentalement. Avec la méthode de regroupement global, toutes les segmentations d'émissions sont considérées simultanément, et l'algorithme de classification travaille avec l'ensemble des classes de locuteurs déterminées par la SRL d'émissions sur chacun des enregistrements qui compose une collection. Avec la méthode de regroupement incrémental, les enregistrements sont traités (ajoutés) itérativement. Le regroupement supplémentaire s'opère lors de chaque itération entre deux segmentations : la segmentation de collections provenant du regroupement de l'itération précédente, et la segmentation d'émission correspondant à l'enregistrement en cours de traitement. Pour une itération donnée, l'algorithme de classification ne travaille qu'avec les classes de locuteurs correspondant à l'actuelle segmentation de collections, plus celles de la segmentation d'émission en cours de traitement.

Les algorithmes de regroupement expérimentés pour effectuer l'étape de regroupement supplémentaire, qu'il s'agisse d'un regroupement global ou incrémental, sont le regroupement agglomératif hiérarchique (HAC) et le regroupement combinatoire ILP. L'approche HAC à l'état de l'art en SRL d'émissions, qui repose sur une modélisation en locuteurs GMM et des scores de vraisemblance croisés (CLR), ne permet pas d'aboutir à des résultats en un temps raisonnable pour le traitement des collections (un constat similaire peut être établi avec l'approche ILP à l'état de l'art en SRL d'émissions, qui repose sur la distance de *Mahalanobis*). Nous avons proposé d'adapter pour le traitement des collections les approches de classification à l'état de l'art en SRL d'émissions. Nous avons, pour ce faire, emprunté des approches issues du domaine de la reconnaissance du locuteur : l'approche de modélisation du locuteur

i-vector, et l'approche de scoring entre les modèles PLDA.

Nous avons présenté une reformulation de la méthode de regroupement ILP, permettant d'alléger le problème en termes de contraintes et de variables. Nous avons également présenté une approche de simplification permettant de réduire un problème de regroupement complexe en plusieurs sous-problèmes indépendants. Cette approche de simplification repose sur la décomposition en composantes connexes d'un graphe non orienté complexe où, les sommets représentent les classes de locuteurs, et les arrêtes, les scores de vraisemblance entre les modèles. La plupart des composantes connexes, 99% en moyenne, correspondent à des problèmes de classification triviaux à résoudre. Seuls les 1% de composantes connexes restantes doivent être résolus par un algorithme de regroupement. Cette approche de simplification, où les composantes connexes complexes sont résolues par ILP ($CC{+}ILP_{PLDA}$) ou HAC ($CC{+}HAC_{PLDA}$), permet d'obtenir un léger gain au niveau des $\text{DER}_{\text{de collections}}$.

Nous avons, finalement, expérimenté deux stratégies pour le regroupement au niveau collection, afin de déterminer s'il est judicieux de laisser la liberté aux approches de regrouper des classes provenant d'un même enregistrement (ce qui correspond à remettre en question la segmentation proposée par le système de SRL d'émissions).

7.1.1 Regroupement global et incrémental

Les approches de regroupement global et incrémental ont été appliquées avec succès pour le traitement des collections. L'approche de regroupement global permet de traiter les collections plus rapidement que l'approche incrémentale, mais les résultats obtenus en termes de $\text{DER}_{\text{de collections}}$, où la détection des locuteurs récurrents est prise en compte lors de l'évaluation, sont légèrement plus mauvais.

Les résultats obtenus en termes de $\text{DER}_{\text{de collections}}$ sont très proches entres les méthodes de classification $CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$, quelles que soient les collections et approches de regroupement (global ou incrémental) étudiées. La méthode de classification $CC{+}HAC_{PLDA}$ permet néanmoins d'atteindre des $\text{DER}_{\text{de collections}}$ légèrement meilleurs, et permet une meilleure détection des locuteurs récurrents, que la méthode de classification $CC{+}ILP_{PLDA}$, avec l'approche de regroupement global. On observe la tendance inverse avec l'approche de regroupement incrémental, où les meilleurs $\text{DER}_{\text{de collections}}$ sont atteints par la méthode de classification $CC{+}ILP_{PLDA}$. En termes de $\text{DER}_{\text{d'émissions}}$, les deux méthodes de classification se valent, quelles que soient les collections et approches de regroupement étudiées.

En termes de locuteurs récurrents correctement détectés par les méthodes, l'approche de regroupement incrémental ne permet pas d'en détecter autant que l'ap-

proche de regroupement global (la différence est cependant minime). En revanche, l'approche de regroupement incrémental fait moins d'erreurs quant à la détection des locuteurs récurrents : parmi les locuteurs récurrents détectés, la proportion qui correspond effectivement à locuteurs récurrents d'après les segmentations de référence est plus élevée qu'avec l'approche de regroupement global.

L'intérêt de l'approche incrémentale réside essentiellement dans le traitement des collections dynamiques, dont le volume est susceptible d'augmenter au cours du temps : alors qu'il serait nécessaire d'effectuer un regroupement global sur l'ensemble des données si un enregistrement venait à compléter une collection déjà traitée, seule une itération supplémentaire de l'approche incrémentale serait nécessaire. Or, la durée de la dernière itération d'un regroupement incrémental ne représente que 41%, en moyenne, de la durée d'un regroupement global intégral.

7.1.2 Collections d'émissions et collections temporelles

La répartition des enregistrements à notre disposition en collections d'émissions et collections temporelles a été réalisée afin d'observer le comportement des méthodes de classification et déterminer si une configuration en particulier était plus propice au traitement des collections. Les résultats obtenus en termes de $\text{DER}_{\text{de collections}}$ obtenus par les deux approches de regroupement (global et incrémental), et les deux méthodes de classification ($CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$), donnent des résultats similaires compris entre 20% et 25% en moyenne, toutes collections confondues. Il est donc difficile de statuer quant au découpage en collection permettant d'optimiser la SRL de collections. Nous avons toutefois constaté que :

- Plus le volume de la collection augmente, plus la tâche de SRL de collections est difficile : les $\text{DER}_{\text{de collections}}$ augmentent, et la proportion de locuteurs récurrents correctement détectés diminue. En revanche, les seuils optimaux pour les méthodes de classification semblent se stabiliser.
- L'approche de regroupement global, avec la méthode de classification $CC{+}HAC_{PLDA}$, est la plus adaptée pour le traitement des collections temporelles. Comparé aux résultats obtenus sur les différentes collections d'émission, le gain en termes de $\text{DER}_{\text{de collections}}$ est de 2,81%, et le gain sur la proportion de locuteurs récurrents correctement détectés est de 7%. Ces résultats sont toutefois à nuancer, car le volume de nos collections temporelles est relativement faible.
- L'approche de regroupement incrémental avec la méthode de classification

CC+ILP$_{PLDA}$ permet d'atteindre des DER$_{\text{de collections}}$ sur les collections d'émissions inférieurs à ceux obtenus par regroupement global (le gain est de 2,37%).

Le choix que nous avons fait pour le traitement des collections d'émissions de niveau *Programme*, où les mêmes seuils β et δ ont été appliqués sur les différentes collections, n'est pas le plus judicieux. La valeur de ces seuils a été choisie de manière à minimiser le DER$_{\text{de collections}}$ moyen, cependant, chacune de ces collections présente des caractéristiques différentes, et donc, des seuils optimaux qui leur sont propres.

7.2. Limites et perspectives

Les approches proposées et expérimentées pour le traitement des collections ont globalement donné des résultats satisfaisants. Des questions se posent cependant quant à la suite des travaux présentés. Nous avons procédé au traitement de collections dont le volume représente jusqu'à 178 heures d'audio. C'est déjà beaucoup, au regard des études déjà réalisées sur cette problématique, mais insignifiant, au regard des volumes qu'il serait souhaitable de traiter. Un problème se pose quant au développement d'approches robustes sur des collections vraiment très volumineuses. Nous ne disposons pas de suffisamment de données annotées à des fins d'évaluations. En passant de la SRL d'émissions à la SRL de collections, nous avons été confrontés à des problèmes liés à l'implémentation de nos approches (en particulier, des problèmes de durées d'exécution). Nous n'aurions pas pris conscience de ces problèmes sans ce changement d'échelle, et nous risquons d'être confrontés à des problèmes similaires si le volume des collections à traiter augmente sensiblement.

Malgré l'approche de simplification permettant la décomposition du problème en sous-problèmes indépendants, la méthode de regroupement *CC+ILP*$_{PLDA}$ échoue si le problème à traiter est trop conséquent. En effet, l'outil de résolution ne semble pas capable de déterminer une solution optimale si de nombreuses classes (plus de 5000) sont impliquées dans un regroupement ILP. Ce problème a été évité avec les collections que nous avons présentées, mais se posera inévitablement à partir du moment où des collections plus volumineuses devront être traitées. Une solution envisageable, hormis expérimenter un nouvel outil de résolution plus performant, serait d'implémenter notre propre outil de résolution. Les outils actuels reposent sur l'algorithme *Branch & Bound*, qui permet de déterminer la solution optimale d'un problème discret en temps non-polynomial. Nous pourrions envisager une approche de résolution moins générale, reposant tout de même sur un algorithme de *backtracking*, pour déterminer plus efficacement une solution s'apparentant à la solution optimale

en nous repliant, par exemple, sur un algorithme polynomial lorsque la durée du calcul dépasse un certain seuil.

L'étape la plus longue dans la chaîne de traitement correspond à l'estimation des scores PLDA entre les modèles de locuteurs. Cette étape est réalisée à l'aide de l'outil *IvTest* de la suite d'outils pour la reconnaissance du locuteur *Alize*. La durée nécessaire à l'estimation des scores pourrait cependant être amoindrie moyennant certaines optimisations directement dans l'outil *IvTest*. Parmi les perfectionnements possibles, nous pouvons citer le *multi-threading* et la gestion interne des matrices. Une autre optimisation, cette fois liée à l'implémentation de nos propres outils, serait de considérer la propriété de symétrie des distances afin de réduire le nombre de scores à estimer (étant donné deux modèles de locuteurs i et j, $S_{PLDA}(i,j) = S_{PLDA}(j,i)$). Nous n'avons malheureusement pris conscience de l'importance de cette optimisation qu'en voulant traiter les collections les plus volumineuses, donc trop tard (nos outils s'attendent à lire des matrices de scores symétriques).

D'une manière plus générale, il serait bon de concevoir un outil vraiment dédié à la SRL de collections, dans un langage de programmation performant, plutôt que de recourir à des scripts faisant appel à des programmes externes. Nous pourrions ainsi minimiser la quantité de ressources nécessaires. À titre d'exemple, nous sommes actuellement contraints d'écrire le contenu des matrices de scores PLDA dans un fichier, de manière à le transmettre aux outils réalisant la classification. Avec l'approche de regroupement global et la collection d'émission *Thématique*, ce fichier de scores PLDA représente un espace disque d'environ 12Go. L'écriture et la lecture de ce fichier sont des opérations coûteuses en temps, d'autant plus si le support de stockage n'est pas directement accessible depuis la machine où sont exécutés les scripts pour la SRL de collections (le réseau pouvant être encombré si de nombreuses opérations distantes sont réalisées en même temps).

7.2.1 L'approche de regroupement incrémental

Le problème évoqué précédemment quant à la durée excessive de l'étape d'estimation des scores PLDA s'est révélé problématique pour l'approche de regroupement incrémental. Cette étape est inévitablement effectuée lors de chaque itération, et le nombre de scores à estimer augmente au fur et à mesure des itérations. Par conséquent, la durée nécessaire pour estimer ces scores, qui est presque insignifiante lors des premières itérations, devient vraiment conséquente à mesure que les itérations se succèdent. Nous avons proposé une approche permettant d'économiser du temps, lors de chaque itération, afin de minimiser la durée totale du regroupement incrémental. Nous utilisons le modèle de locuteur correspondant à la classe la plus centrale d'un

regroupement pour représenter la nouvelle classe obtenue, plutôt que de construire un nouveau modèle à partir des données des classes regroupées (*recyclage* des modèles de locuteurs). Cette approche dégrade très légèrement les résultats en termes de $\text{DER}_{\text{de collections}}$ (la différence est inférieure à 1%, en moyenne, sur les collections d'émissions de niveau *Programme*) mais permet une réduction moyenne de la durée totale du regroupement incrémental d'environ 77%.

Nous envisageons, dans de futurs travaux, d'examiner l'influence de l'ordre de traitement des enregistrements d'une collection. Nous avons opté pour l'ordre chronologique des enregistrements, en revanche, nous n'avons pas déterminé son impact sur les résultats du procédé de regroupement incrémental.

L'intérêt de l'approche de regroupement incrémental réside essentiellement dans le traitement des collections pour lesquelles des enregistrements seraient fréquemment ajoutés, or, des améliorations sont encore nécessaires pour procéder au traitement de collections vraiment très volumineuses. Nous avons dû faire face à différents problèmes avec les 178 heures d'audio que représente la plus volumineuse de nos collections, comment allons-nous nous y prendre pour traiter un flux de données quotidien comme les 300 heures de vidéo publiées chaque minute sur YouTube, ou les millions d'heures stockés par l'INA ? Dans l'état actuel de nos recherches, le traitement de telles collections serait beaucoup trop long, même en optimisant au mieux nos approches de regroupement et méthodes de classification.

7.2.2 L'analyse des résultats de classification

La métrique $\text{DER}_{\text{de collections}}$ permet d'évaluer les performances de la SRL de collections. Toutefois, ce taux d'erreur ne donne qu'une idée générale de la qualité des segmentations produites : on ne dispose pas d'informations concrètes sur la détection des locuteurs récurrents, et il est difficile, étant donné la quantité de données à analyser, de déterminer avec précision la raison des erreurs de regroupement. S'il est assez simple d'analyser le comportement des approches de regroupement en SRL d'émissions, en visualisant, par exemple, le fichier audiovisuel traité (des outils d'analyse existent), c'est beaucoup plus complexe sur une collection, d'autant plus si le nombre d'enregistrements impliqué est élevé.

La métrique DER ne nous a pas semblé suffisamment discriminante pour déterminer laquelle des méthodes de classification proposées est la plus pertinente : les $\text{DER}_{\text{de collections}}$ et $\text{DER}_{\text{d'émissions}}$ sont trop peu dissemblables entre les méthodes de classification (quelles que soient la collection et l'approche de regroupement étudiées). Le choix de la méthode de classification à adopter devrait plutôt se faire en

fonction des contraintes algorithmique et technique, ainsi que du domaine d'application visé.

Annoter de nombreuses données à des fins d'évaluation est une solution viable pour la recherche. En revanche, comment estimer les performances de la SRL de collections dans un contexte applicatif industriel ? Leurs données ne seront pas annotées, d'autres solutions doivent être envisagées : utiliser des modèles de locuteurs *a priori* ? Utiliser des systèmes semi-supervisés où un annotateur humain guiderait le procédé de SRL ? La question reste en suspens, et les réponses ne sont, dans l'état actuel des choses, pas nombreuses.

Quatrième partie

Annexes

ANNEXE A

Regroupement global : analyse intermédiaire

Nous présentons dans cette annexe les analyses intermédiaires effectuées sur les différentes collections traitées avec l'architecture de regroupement global. Les approches de regroupement globales ayant été comparées sont celles basées sur la décomposition en composantes connexes et les scores de vraisemblance PLDA : $CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$. La modélisation en locuteurs est identique pour les deux approches de regroupement, il s'agit de modèles i-vector de dimension 300 normalisés par une itération de l'approche SNN (Spherical Nuisance Normalization). La SRL d'émission, effectuée indépendamment sur chaque enregistrement au niveau *émission* de l'architecture, correspond à celle présentée dans la section 5.2.2 : un système de SRL d'émissions à l'état de l'art où les classes issues du regroupement BIC ($\lambda = 3$) sont modélisées par des i-vectors de dimension 300 et regroupées par l'approche ILP_{PLDA} avec un seuil $\delta = 20$. Les seuils choisis pour les approches de regroupement global $CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$ ont été sélectionnés en fonction des « règles » énoncées en partie 5.2.3, où concernant l'approche CC+ILP, $\beta = \delta$, et concernant l'approche CC+HAC, le seuil β doit être identique à celui de l'approche CC+ILP. Le critère d'optimisation pour la sélection des seuils β et δ, étant donné ces deux règles, repose sur le $\text{DER}_{\text{de collections}}$ moyen minimum.

Cette annexe s'organise en deux sections. Nous présentons dans un premier temps une analyse menée sur les collections d'émissions (chacun des niveaux *programme*, *organisme (chaîne)* et *thématique* faisant l'objet d'une sous-section dédiée). Nous présentons ensuite une analyse similaire sur les 8 collections *Temporelles*. Ce sont les différentes informations présentées dans cette annexe qui ont servi à établir les observations formulées dans la partie *5.4. Architecture de regroupement global : évaluation* du chapitre 5. Nous présentons en détail pour chacune des collections étudiées :

- Le nombre total de locuteurs et le nombre de locuteurs récurrents dans les segmentations de référence.
- Les taux d'erreur $\text{DER}_{\text{de collections}}$ et $\text{DER}_{\text{d'émissions}}$ obtenus aux niveaux *émission* (ILP_{PLDA}) et *collection* ($CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$) de l'architecture de regroupement global.
- Le nombre total de locuteurs (classes) et le nombre de locuteurs récurrents (classes formées d'au moins deux segments provenant d'enregistrements différents) résultant du regroupement global, ainsi que le nombre de locuteurs récurrent (classes) correspondant à des locuteurs effectivement récurrents d'après les segmentations de référence.

Le nombre de locuteurs récurrents détectés par le système correspondant effectivement à des locuteurs récurrents d'après les segmentations de référence peut être facilement déterminé grâce à l'outil d'évaluation, qui permet de générer la correspondance entre les étiquettes des segmentations de référence et celles fournies par le système. Nous considérons qu'un locuteur récurrent est correctement détecté par le système si l'étiquette de la classe qui le représente est associée à des segments provenant d'au moins deux enregistrements différents, et bien sûr, si le locuteur correspondant dans les segmentations de référence est effectivement récurrent. Nous fournissons également, à titre indicatif, les durées relatives à l'accomplissement des procédés de regroupements mis en œuvre. Nous présentons ainsi, pour chacune des collections étudiées :

- La durée totale de la collection en termes de modalité audio et le nombre de classes de locuteurs impliquées dans le problème de regroupement global.
- La durée effective pour le calcul des scores PLDA entre les différentes paires de modèles de locuteurs (étant donné les modèles de locuteurs).
- Le nombre de composantes connexes complexes déterminées par l'approche de décomposition et le nombre minimum, moyen et maximum de classes par composante connexe complexe.
- La durée effective pour les regroupements $CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$ (qui incluent la décomposition en composantes connexes et le regroupement effectué sur les composantes connexes complexes).

Le calcul des durées effectives a été réalisé dans les mêmes conditions, et sur la même machine, pour chacune des collections étudiées.

A.1. Collections d'émissions

A.1.1 Niveau Programme

Les collections d'émissions du niveau *programme* sont celles sur lesquelles ont été expérimentées les différentes approches présentées dans le chapitre 5, il ne s'agit donc que d'un report des résultats préalablement présentés. En revanche, le lecteur aura cette fois-ci le plaisir d'apprécier les médiocres résultats obtenus sur la collection *Planète Showbiz*... Nous présentons également les DER moyens obtenus avec et sans prise en compte de cette collection problématique. Il est toutefois bon de rappeler que les seuils β et δ ont été sélectionnés de manière à minimiser le $\text{DER}_{\text{de collection}}$ moyen sur l'ensemble des collections de niveau *Programme*, *Planète Showbiz* mise à part. Il s'agit donc des seuils qui, en moyenne, permettent d'obtenir les meilleurs $\text{DER}_{\text{de collection}}$ (*Planète Showbiz* toujours mise à part).

	n^{bre} locuteurs références	**ILP_{PLDA} niveau émission**	**CC. + ILP_{PLDA} niveau collection**	**CC. + HAC_{PLDA} niveau collection**
Seuil β		-	-30	-30
Seuil δ		20	-30	30
BFM Story	[556 ; **83**]	44,75% (10,70%) [853 ; 0]	15,03% (10,70%) [691 ; 58 ; **42**]	14,72% (10,70%) [691 ; 58 ; **43**]
Planète Showbiz	[771 ; **75**]	78,76% (37,41%) [890 ; 0]	72,74% (37,41%) [693 ; 37 ; **6**]	69,06% (37,41%) [690 ; 40 ; **9**]
Ça vous Regarde	[173 ; **11**]	35,73% (18,51%) [212 ; 0]	21,82% (18,51%) [184 ; 7 ; **5**]	21,82% (18,51%) [184 ; 7 ; **5**]
Entre les Lignes	[18 ; **9**]	85,77% (15,01%) [158 ; 0]	15,34% (15,01%) [39 ; 10 ; **8**]	15,34% (15,01%) [39 ; 10 ; **8**]
LCP Info	[317 ; **96**]	61,42% (9,60%) [711 ; 0]	21,30% (9,60%) [538 ; 42 ; **32**]	20,20% (9,51%) [532 ; 40 ; **32**]
Pile et Face	[46 ; **15**]	43,62% (20,05%) [116 ; 0]	23,74% (20,05%) [61 ; 18 ; **14**]	23,74% (20,05%) [61 ; 18 ; **14**]
Top Questions	[119 ; **36**]	57,20% (13,31%) [346 ; 0]	29,24% (13,31%) [193 ; 38 ; **31**]	31,28% (13,31%) [191 ; 37 ; **30**]
Moyenne		**54,83% (14,92%)**	**23,29% (14,92%)**	**23,18% (14,94%)**
Moy. sans Planète Showbiz		**52,97% (13,17%)**	**19,43% (13,17%)**	**19,35% (13,15%)**

Table *A.1 – Résultats obtenus avec l'architecture de regroupement global sur les collections de niveau* programme*, avec les approches de regroupement HAC_{PLDA} et ILP_{PLDA}, avec application de l'approche de décomposition en composantes connexes. Pour chaque approche et chaque collection sont présentés : le $DER_{\text{de collections}}$, avec le $DER_{\text{d'émissions}}$ entre parenthèses, et le nombre de locuteurs détectés par le système (formalisme : [n^{bre} locuteurs total détecté par le système ; n^{bre} locuteurs récurrents détectés ;* **n^{bre} locuteurs détectés effectivement récurrents d'après les références***]).*

Une autre approche aurait pu être de déterminer le couple idéal de seuils β et δ

propre à chacune de ces 7 collections. Cela aurait cependant nui à la comparaison entre les approches de regroupement, d'où notre choix de sélectionner un unique couple de seuils β et δ par approche de regroupement.

Que l'on considère ou non la collection *Planète Showbiz* dans le calcul des DER moyens, les $\text{DER}_{\text{de collections}}$ obtenus par les deux approches de regroupement *CC+HAC* et *CC+ILP* sont très proches (-0,11% en faveur de l'approche *CC+HAC*) (*cf.* tableau A.1). Les $\text{DER}_{\text{d'émissions}}$, calculés sur les segmentations issues du regroupement global, restent quant à eux relativement stables par rapport à ceux obtenus à partir des segmentations d'*émissions* (où chaque enregistrement est traité séparément par le système de SRL d'émissions), avec de légères variations de l'ordre de 0,02% en moyenne.

Les approches de regroupement *CC+ILP* et *CC+HAC* permettent de détecter correctement environ 58%, en moyenne, des locuteurs récurrents d'après les segmentations de référence.

Collection	Durée audio	n^{bre} classes	Durée calcul PLDA	n^{bre} CC. compl.	n^{bre} classes par CC. complexe	Durée CC. + ILP_{PLDA}	Durée CC. + HAC_{PLDA}
BFM Story	49h32m	2845	30m36s	18	4 ; **28** ; 192	53s	38s
Planète Showbiz	38h27m	4224	1h51m	30	4 ; **45** ; 912	1m50s	1m54s
Ça vous Regarde	20h55m	814	48s	10	4 ; **13** ; 40	6s	5s
Entre les Lignes	16h14m	732	34s	5	6 ; **53** ; 146	7s	4s
LCP Info	20h34m	1812	8m21s	15	5 ; **36** ; 168	39s	16s
Pile et Face	19h44m	868	55s	8	5 ; **35** ; 111	8s	5s
Top Questions	12h20m	1000	1m27s	8	4 ; **52** ; 150	10s	6s

Table *A.2 – Durées des approches de regroupement avec l'architecture de regroupement global sur les collections de niveau* programme. *La 2e colonne présente le nombre de classes impliquées dans les problèmes de regroupement; Les 4e et 5e colonnes présentent respectivement le nombre de composantes connexes complexes déterminées par l'approche de décomposition et le nombre (minimum ; moyen ; maximum) de classes par composante connexe complexe.*

La durée nécessaire à l'accomplissement des regroupements *CC+ILP* et *CC+HAC* est presque insignifiante au regard de la durée nécessaire au calcul des scores PLDA entre chacune des paires de modèles de locuteurs impliqués dans le problème de regroupement (*cf.* tableau A.2). On observe d'ailleurs une très forte corrélation entre ces durées et le nombre de classes impliquées, et la durée requise pour estimer les scores de vraisemblance entre les modèles croît très vite en fonction du nombre de classes. Le nombre de composantes connexes complexes augmente également en fonction du nombre total de classes impliquées dans le problème de regroupement, cependant, les sous-problèmes de regroupement correspondant restent triviaux, n'impliquant en moyenne que 13 à 53 classes.

A.1.2 Niveau Organisme

Les collections de niveau *Organisme* (ou, *chaînes de télévision*) sont constituées, pour la collection *BFMTV* de tous les enregistrements des collections de niveau *Programme BFM Story* et *Planète Showbiz*, et pour la collection *LCP*, de tous les enregistrements des collections de niveau *programme Ça vous Regarde*, *Entre les Lignes*, *LCP Info*, *Pile et Face* et *Top Questions*. Les collections *BFMTV* et *LCP* sont beaucoup plus coûteuses à traiter étant donné le nombre d'enregistrements, et donc de classes de locuteurs, qui les composent. La concaténation des segmentations d'émissions est constituée de 7069 classes de locuteurs pour *BFMTV*, et 5226 classes de locuteurs pour *LCP*.

	n^{bre} locuteurs références	**ILP_{PLDA} niveau émission**	**CC. + ILP_{PLDA} niveau collection**	**CC. + HAC_{PLDA} niveau collection**
Seuil β		-	-20	-20
Seuil δ		20	-20	60
BFMTV	[1309 ; **160**]	50,83% (15,34%) [1743 ; 0]	24,20% (15,34%) [1282 ; 80 ; **45**]	22,01% (15,38%) [1318 ; 103 ; **61**]
LCP	[544 ; **172**]	63,28% (14,62%) [1543 ; 0]	26,35% (14,61%) [942 ; 148 ; **101**]	25,81% (14,72%) [928 ; 139 ; **101**]
Moyenne		**58,10% (14,92%)**	**25,45% (14,91%)**	**24,22% (14,99%)**

Table *A.3 – Résultats obtenus avec l'architecture de regroupement global sur les deux collections de niveau* organisme*, avec les approches de regroupement HAC_{PLDA} et ILP_{PLDA}, avec application de l'approche de décomposition en composantes connexes. Pour chaque approche et chaque collection sont présentés : le $DER_{de\ collections}$, avec le $DER_{d'émissions}$ entre parenthèses, et le nombre de locuteurs détectés par le système (formalisme : [n^{bre} locuteurs total détecté par le système ; n^{bre} locuteurs récurrents détectés ;* **n^{bre} locuteurs détectés effectivement récurrents d'après les références***]).*

Concernant les taux d'erreurs $DER_{de\ collections}$, l'écart se creuse entre les résultats obtenus par les deux approches de regroupement *CC+HAC* et *CC+ILP*. L'approche de regroupement *CC+HAC* permet d'aboutir à un résultat moyen inférieur de 1,23%, en termes de $DER_{de\ collections}$, à celui de l'approche de regroupement *CC+ILP* (cet écart n'était que de 0,11%, en moyenne, sur les collections de niveaux *programme*). Cette constatation s'inverse cependant en ce qui concerne les $DER_{d'émissions}$ moyens.

La proportion de locuteurs récurrents correctement détectés par les méthodes de classification diminue, comparée à celle constatée sur les collections d'émissions de niveau *Programme*, ce qui n'est pas tellement surprenant étant donné l'augmentation des $DER_{de\ collections}$. En revanche, cette proportion qui était similaire entre les deux méthodes de classification, avec les collections d'émissions de niveau *Programme*, est cette fois-ci différente : l'approche de regroupement *CC+ILP* ne détecte correctement que 43,4%, en moyenne, des locuteurs récurrents d'après les segmentations de référence. L'approche de regroupement *CC+HAC* en détecte 48,4%.

Au-delà de ces constatations sur les résultats obtenus, nous souhaitons partager une observation à propos de la collection *BFMTV*. Nous avons rencontré un problème avec le regroupement global *CC+ILP*. Afin de déterminer les seuils β et δ permettant d'atteindre les taux d'erreurs les plus faibles, nous avons testé différentes combinaisons de valeurs en faisant varier les seuils de -100 à 100 avec un pas de 10. Le regroupement *CC+ILP* a échoué avec $\beta = \delta = 70$ et $\beta = \delta = 80$. Il s'agit du problème déjà évoqué dans la partie 5.2.2, où nous avons observé une valeur *limite* à partir de laquelle les problèmes soumis à l'outil de résolution sont trop complexes (volumineux) pour être résolus dans un laps de temps raisonnable. La collection *BFMTV* est constituée de 7069 classes de locuteurs. Avec une valeur de β égale à 80, la composante connexe complexe posant problème implique 6225 classes de locuteurs, soit environ 88% de l'ensemble des classes. Avec une valeur de β égale à 70, la composante connexe complexe posant problème implique 5662 classes de locuteurs (environ 80% de l'ensemble des classes de la collection). Il semble donc que cette *limite* quant à la complexité des problèmes soumis à l'outil de résolution, que nous avions repoussé par la décomposition en composantes connexes, refait surface dès lors que les collections traitées commencent à prendre du volume. Il convient cependant de relativiser, car les seuils problématiques sont loin du seuil optimal (-20).

Collection	Durée audio	n^{bre} classes	Durée calcul PLDA	n^{bre} CC. complexes	n^{bre} classes par CC. complexe	Durée CC. + ILP_{PLDA}	Durée CC. + HAC_{PLDA}
BFMTV	88h	7069	8h37m	58	4 ; **21** ; 132	3m32s	6m13s
LCP	89h48	5226	3h18m	36	4 ; **37** ; 246	2m18s	3m42s

Table *A.4 – Durées des approches de regroupement avec l'architecture de regroupement global sur les collections de niveau* organisme. *La 2e colonne présente le nombre de classes impliquées dans les problèmes de regroupement; Les 4e et 5e colonnes présentent respectivement le nombre de composantes connexes complexes déterminées par l'approche de décomposition et le nombre (minimum ; moyen ; maximum) de classes par composante connexe complexe.*

Les durées relevées pour l'accomplissement des regroupements des collections de niveau *Organisme* sont en adéquations avec celles observées pour le traitement des collections de niveau *Programme* (*cf.* tableau A.4). La durée des regroupements est négligeable comparé à la durée nécessaire pour extraire les modèles de locuteurs et celle nécessaire à l'estimation des scores PLDA entre chaque paire de modèles de locuteurs impliqués dans le problème de regroupement, ce qui n'est pas surprenant étant donné le faible coût de l'étape permettant la décomposition en composantes connexes, et le nombre restreint de composantes connexes complexes. Étant donné le nombre total de classes impliquées dans le problème de regroupement (7069 pour la collection *BFMTV* et 5226 pour *LCP*), le nombre de composantes connexes complexes déterminées par l'approche de décomposition (respectivement 58 et 36), et le

nombre moyen de classes impliquées dans les composantes connexes complexes (respectivement 21 et 37), on peut estimer qu'environ 83% de l'ensemble du problème regroupement de la collection *BFMTV* (respectivement 75% pour la collection *LCP*) est résolu par la seule recherche des composantes connexes étoilées.

A.1.3 Niveau Thématique

Notre collection de niveau *Thématique* est composée de l'intégralité des enregistrements dont nous disposons. Il s'agit de la collection la plus fournie et la plus diversifiée. Le regroupement du niveau *collection* porte sur 12295 classes de locuteurs. Ici aussi, le regroupement global *CC+ILP* a échoué pour le traitement de certaines composantes connexes complexes, pour certaines combinaisons de seuils (en particulier, avec $\beta = \delta >= 60$). Que l'approche de regroupement global *CC+ILP* échoue en fonction de la valeur des seuils est gênant, mais pas préjudiciable étant donné que nous sommes loin de la valeur optimale, qui est de -20.

	n^{bre} locuteurs références	**ILP_{PLDA} niveau émission**	**CC. + ILP_{PLDA} niveau collection**	**CC. + HAC_{PLDA} niveau collection**
Seuil β		-	-20	-20
Seuil δ		20	-20	60
BFMTV + LCP	[1787 ; **333**]	59,23% (14,92%) [3286 ; 0]	26,24% (14,91%) [2266 ; 257; **151**]	24,60% (14,98%) [2233 ; 246 ; **158**]

Table *A.5 – Résultats obtenus avec l'architecture de regroupement global sur la collection de niveau* thématique *(toutes les données disponible), avec les approches de regroupement HAC_{PLDA} et ILP_{PLDA}, avec application de l'approche de décomposition en composantes connexes. Pour chaque approche et chaque collection sont présentés : le $DER_{de\ collections}$, avec le $DER_{d'émissions}$ entre parenthèses, et le nombre de locuteurs détectés par le système (formalisme : [n^{bre} locuteurs total détecté par le système ; n^{bre} locuteurs récurrents détectés ;* **n^{bre} locuteurs détectés effectivement récurrents d'après les références***]).*

L'écart entre les résultats des approches de regroupement *CC+ILP* et *CC+HAC*, en termes de $DER_{de\ collections}$, continue de se creuser : -1,64% en faveur de l'approche de regroupement *CC+HAC*. Autre constatation digne d'intérêt, les combinaisons de seuil β et δ sélectionnées pour minimiser les $DER_{de\ collections}$ avec les approches *CC+ILP* et *CC+HAC* sont les mêmes que celles sélectionnées pour les collections du niveau *Organisme*. Il ne s'agit que de conjectures étant donné le nombre de DER intervenant dans le calcul des moyennes, mais il semble que :

1. L'approche de regroupement *CC+HAC* est plus robuste que l'approche de regroupement *CC+ILP*. En effet, les résultats moyens obtenus sur les collections des différents niveaux sont plus stables.

2. Plus la quantité de données à traiter augmente et se diversifie, plus la combinaison de seuils β et δ *idéale* se stabilise, pour les deux approches de regroupement.

La proportion de locuteurs récurrents correctement détectés par les méthodes de classification est de 45,4% pour la méthode *CC+ILP*, et 47,4% pour la méthode *CC+HAC*. Ces proportions sont du même ordre que celles observées avec les collections d'émissions de niveau *Organisme*.

Collection	Durée audio	n^{bre} classes	Durée calcul PLDA	n^{bre} CC. complexes	n^{bre} classes par CC. complexe	Durée CC. + ILP_{PLDA}	Durée CC. + HAC_{PLDA}
BFMTV + LCP	177h48	12295	50h29m	101	4 ; **27** ; 246	11m9s	29m40s

Table *A.6 – Durées des approches de regroupement avec l'architecture de regroupement global sur la collection de niveau* thématique. *La 2e colonne présente le nombre de classes impliquées dans les problèmes de regroupement; Les 4e et 5e colonnes présentent respectivement le nombre de composantes connexes complexes déterminées par l'approche de décomposition et le nombre (minimum ; moyen ; maximum) de classes par composante connexe complexe.*

Le traitement de la collection de niveau *thématique*, qui englobe l'ensemble des données étudiées jusqu'à présent, est relativement long (*cf.* tableau A.6). La durée des approches de regroupement est toujours négligeable comparé à la durée réelle de la modalité audio et au nombre de classes impliquées, cependant, l'estimation des scores entre les différentes paires de modèles de locuteurs est coûteux (12295^2 calculs). À titre informatif, nous mémorisons les scores PLDA entre chaque paire dans un fichier texte, avec un score par ligne (sous la forme « *modèle_A modèle_B score* »). Avec cette collection de niveau *thématique*, la taille en notre fichier de scores PLDA est d'environ 12,72 Go. Il est composé de 151167025 lignes, et donc, coûteux à parcourir. Bien entendu, ce choix d'implémentation avait été fait pour traiter des émissions où la taille et le stockage n'impactaient en rien les temps de calcul. L'approche de décomposition à conduit à l'établissement de 101 composantes connexes complexes composées, en moyenne, de 27 classes. Au regard du nombre total de classes constituant le problème de regroupement sur la collection *BFMTV+LCP*, on estime qu'environ 78% de ce problème est résolu par la seule recherche des composantes connexes étoilées.

A.2. Collections temporelles

Les collections *temporelles* ont été constituées à partir de l'ensemble des enregistrements dont nous disposons. La répartition des enregistrements a été réalisée en

fonction des « irrégularités » observées dans la fréquence d'acquisition des enregistrements. Nous avons, dans la mesure du possible, regroupé les enregistrements consécutifs en considérant l'ordre chronologique de diffusion. Étant donné cet ordre, nous avons constitué une nouvelle collection dès lors que deux enregistrements consécutifs étaient séparés par un minimum de 15 jours (*cf.* figure 4.2). La durée des collections *temporelles* est très hétérogène, elle varie entre environ 6 et 60 heures d'audio selon la collection étudiée.

	n^{bre} locuteurs références	**ILP_{PLDA} niveau émission**	**CC. + ILP_{PLDA} niveau collection**	**CC. + HAC_{PLDA} niveau collection**
Seuil β		-	10	10
Seuil δ		20	10	40
Temporelle n°1	[58 ; **10**]	27,21% (11,71%) [89 ; 0]	15,01% (13,59%) [68 ; 12 ; **10**]	15,01% (13,59%) [68 ; 12 ; **10**]
Temporelle n°2	[217 ; **26**]	28,15% (13,19%) [263 ; 0]	19,88% (13,19%) [219 ; 28 ; **15**]	21,55% (14,75%) [219 ; 25 ; **15**]
Temporelle n°3	[140 ; **28**]	33,29% (12,91%) [170 ; 0]	15,61% (12,81%) [139 ; 23 ; **20**]	15,61% (12,81%) [139 ; 23 ; **20**]
Temporelle n°4	[669 ; **108**]	49,63% (13,79%) [986 ; 0]	20,39% (13,21%) [658 ; 89 ; **59**]	20,14% (13,31%) [683 ; 102 ; **62**]
Temporelle n°5	[247 ; **51**]	42,87% (13,31%) [402 ; 0]	20,94% (13,48%) [298 ; 40 ; **26**]	18,82% (13,45%) [303 ; 39 ; **27**]
Temporelle n°6	[205 ; **30**]	65,38% (24,66%) [309 ; 0]	32,60% (24,57%) [228 ; 27 ; **17**]	31,88% (24,49%) [236 ; 29 ; **17**]
Temporelle n°7	[262 ; **37**]	38,03% (11,88%) [329 ; 0]	14,46% (12,06%) [271 ; 30 ; **23**]	14,30% (11,88%) [273 ; 31 ; **23**]
Temporelle n°8	[461 ; **64**]	50,23% (19,14%) [682 ; 0]	24,63% (19,24%) [478 ; 66 ; **43**]	24,21% (19,37%) [495 ; 63 ; **41**]
Moyenne		**44,62% (14,99%)**	**20,66% (14,97%)**	**20,37% (15,13%)**

Table *A.7 – Résultats obtenus avec l'architecture de regroupement global sur les collections* temporelles*, avec les approches de regroupement HAC_{PLDA} et ILP_{PLDA}, avec application de l'approche de décomposition en composantes connexes. Pour chaque approche et chaque collection sont présentés : le $DER_{de\ collections}$, avec le $DER_{d'émissions}$ entre parenthèses, et le nombre de locuteurs détectés par le système (formalisme : [n^{bre} locuteurs total détecté par le système ; n^{bre} locuteurs récurrents détectés ;* **n^{bre} locuteurs détectés effectivement récurrents d'après les références***]).*

Le choix des seuils β et δ, déterminé expérimentalement, a porté sur le critère de $DER_{de\ collections}$ moyen minimum (toujours sous la contrainte que $\beta = \delta$ pour l'approche de regroupement *CC+ILP*, et un seuil β identique pour l'approche de regroupement *CC+HAC*, afin de rester cohérent dans la comparaison des deux approches de regroupement). Ces seuils ont ensuite été appliqués tels quels sur les 8 collections *temporelles* étudiées. Le comportement entre les deux approches de regroupement global *CC+ILP* et *CC+HAC* est semblable à celui observé sur les collections *d'émissions* : les résultats moyens obtenus sont très proches, avec une légère supériorité

pour le regroupement *CC+HAC* en ce qui concerne le $\text{DER}_{\text{de collections}}$ (20,37% contre 20,66%). Les résultats en termes de $\text{DER}_{\text{d'émissions}}$ sont également très proches, c'est cependant l'approche de regroupement *CC+ILP* qui permet d'atteindre le plus faible résultat (15,13% contre 14,97%).

La proportion de locuteurs récurrents correctement détectés par les méthodes de classification, d'après les segmentations de référence, est d'environ 65%. Il s'agit de la meilleure proportion observée avec l'architecture de regroupement global, toutes collections confondues. Cette particularité peut s'expliquer par le fait que toutes les collections temporelles sont constituées d'enregistrements de provenance hétérogène. Les seuils β et δ, sélectionnés pour optimiser le $\text{DER}_{\text{de collections}}$ moyen, sont donc probablement plus robustes.

Il semble donc que le découpage effectué pour constituer les collections temporelles ait une influence positive à la fois sur le $\text{DER}_{\text{de collections}}$ et sur le nombre de locuteurs récurrents correctement détectés par le système, comparé au découpage en collections d'émissions. En effet, les $\text{DER}_{\text{de collections}}$ sont sensiblement inférieurs à ceux constatés sur les différentes collections d'émissions, avec un gain absolu moyen de 2,63% pour l'approche de regroupement *CC+ILP* et 2,81% pour l'approche de regroupement *CC+HAC* (par rapport aux meilleurs résultats obtenus sur les collections d'émissions). Par rapport au nombre de locuteurs récurrents correctement détectés, le gain absolu moyen est d'environ 7% pour les deux approches de regroupement global (toujours par rapport aux meilleurs résultats constatés sur les collections d'émissions).

Collection	Durée audio	n^{bre} classes	Durée calcul PLDA	n^{bre} CC. complexes	n^{bre} classes par CC. complexe	Durée CC. + ILP_{PLDA}	Durée CC. + HAC_{PLDA}
Temporelle n°1	6h11m	266	2s	1	5 ; **5** ; 5	< 1s	< 1s
Temporelle n°2	15h42m	906	1m7s	4	5 ; **10** ; 21	4s	4s
Temporelle n°3	10h20m	630	23s	0	0 ; **0** ; 0	1s	2s
Temporelle n°4	39h44m	3010	38m50s	25	4 ; **18** ; 69	45s	32s
Temporelle n°5	18h02m	1411	4m7s	11	4 ; **10** ; 26	11s	9s
Temporelle n°6	9h29m	927	1m14s	5	10 ; **14** ; 24	7s	4s
Temporelle n°7	12h36m	683	29s	3	4 ; **9** ; 15	2s	1s
Temporelle n°8	61h55m	4237	1h48m	40	4 ; **14** ; 76	1m23s	1m16s

Table *A.8 – Durées des approches de regroupement avec l'architecture de regroupement global sur les collections* temporelles. *La 2e colonne présente le nombre de classes impliquées dans les problèmes de regroupement; Les 4e et 5e colonnes présentent respectivement le nombre de composantes connexes complexes déterminées par l'approche de décomposition et le nombre (minimum ; moyen ; maximum) de classes par composante connexe complexe.*

En matière de durées (*cf.* tableau A.8), les constats sont similaires à ceux déjà établis sur les collections *d'émissions*. La durée nécessaire à l'accomplissement des

approches de regroupement *CC+ILP* et *CC+HAC* est presque insignifiante comparé à la durée requise pour calculer le score PLDA entre les paires de modèles de locuteurs. L'approche de décomposition en composantes connexes montre encore une fois sont intérêt, notamment avec la collection *Temporelle n°3* où aucune composante connexe complexe n'a été mise en évidence. Le problème de regroupement correspondant à la collection *Temporelle n°3* a donc été intégralement résolu par la recherche de composantes connexes en *étoilées*.

ANNEXE B

Regroupement incrémental : analyse intermédiaire

Nous présentons dans cette annexe une série d'analyses intermédiaires effectuées sur les collections traitées avec l'architecture de regroupement incrémental. Les collections concernées sont les collections d'émissions de niveau *Programme* et *Thématique*, ainsi que les huit collections *Temporelles*.

Les approches de classification comparées pour le regroupement incrémental reposent sur la décomposition en composantes connexes et les scores de vraisemblance PLDA : $CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$. La modélisation en locuteurs, pour les deux approches de classification, reste inchangée : il s'agit de modèles i-vector de dimension 300 extraits à l'aide du modèle du monde à 1024 composantes gaussiennes et normalisés par SNN (une seule itération). Les segmentations d'émissions de chaque enregistrement correspondent à celles présentées dans la section 5.2.2 : les classes issues du regroupement BIC ($\lambda = 3$) sont représentées par des modèles i-vectors semblables, de dimension 300, et regroupées par l'approche ILP_{PLDA} avec un seuil $\delta = 20$. Les seuils sélectionnés pour les approches de regroupement incrémental $CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$, pour chacune des collections étudiées, sont ceux déterminés pour expérimenter l'approche de regroupement global. La comparaison entre les résultats obtenus avec les approches de regroupement global et incrémental ne porte donc que sur la manière d'envisager la SRL de collections sur les enregistrements des différentes collections.

Cette annexe s'organise de la même manière que la précédente, en deux sections. La première concerne les collections d'émissions de niveaux *Programme*, et la seconde, les collections temporelles. Nous présentons pour chacune des collections étudiées :

- Le nombre total de locuteurs et le nombre de locuteurs récurrents dans les segmentations de référence.
- Les taux d'erreur $DER_{\text{de collections}}$ et $DER_{\text{d'émissions}}$ obtenus aux niveaux *émission* (ILP_{PLDA}) et *collection* ($CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$) de l'architecture de regroupement incrémental.
- Le nombre total de locuteurs (classes) et le nombre de locuteurs récurrents (classes formées d'au moins deux segments provenant d'enregistrements différents) résultant du regroupement incrémental, ainsi que le nombre de locuteurs récurrent (classes) correspondant à des locuteurs effectivement récurrents d'après les segmentations de référence.

Le nombre de locuteurs récurrents détectés par le système correspondant effectivement à des locuteurs récurrents d'après les segmentations de référence est déterminé grâce à la correspondance entre les étiquettes des segmentations de référence et celles fournies par le système (cette correspondance est établie par l'outil d'évaluation). Nous avons considéré qu'un locuteur récurrent est correctement détecté par le système si l'étiquette de la classe qui le représente est associée à des segments provenant d'au moins deux enregistrements différents, et si le locuteur correspondant dans les segmentations de référence est effectivement récurrent.

Nous présentons également une comparaison entre les résultats obtenus avec et sans application du procédé de *Recyclage*. Ce procédé, présenté en détail dans la partie 6.2, consiste à minimiser le nombre de modèles i-vector à extraire entre chaque itération du regroupement incrémental : pour une itération donnée, le modèle de locuteur censé représenter une classe regroupée durant l'itération précédente correspond en fait au modèle i-vector de la classe « au centre » de ce regroupement.

B.1. Collections d'émissions

B.1.1 Niveau Programme

Contrairement aux observations énoncées par Tran et al. [2011], nos résultats en termes de $DER_{\text{de collections}}$, obtenus sur les collections d'émissions du niveau *Programme* avec l'approche de regroupement incrémental, sont, en moyenne, meilleurs que ceux obtenus avec l'approche de regroupement global. Nous présentons, dans le tableau B.1, les résultats obtenus en termes de DER ainsi que des informations sur le nombre de locuteurs récurrents détectés par les approches de regroupement

incrémental *CC+ILP*$_{PLDA}$ et *CC+HAC*$_{PLDA}$ pour chacune des collections de niveau *Programme*.

	n^{bre} locuteurs références	**ILP$_{PLDA}$ niveau émission**	**CC. + ILP$_{PLDA}$ niveau collection**	**CC. + HAC$_{PLDA}$ niveau collection**
Seuil β		-	-30	-30
Seuil δ		20	-30	30
BFM Story	[556 ; **83**]	44,75% (10,70%) [853 ; 0]	14,32% (10,35%) [678 ; 54 ; **43**]	15,65% (11,09%) [678 ; 53 ; **42**]
Planète Showbiz	[771 ; **75**]	78,76% (37,41%) [890 ; 0]	47,35% (36,49%) [651 ; 16 ; **7**]	47,35% (37,41%) [651 ; 16 ; **7**]
Ça vous Regarde	[173 ; **11**]	35,73% (18,51%) [212 ; 0]	21,90% (18,51%) [183 ; 8 ; **5**]	21,90% (18,51%) [183 ; 8 ; **5**]
Entre les Lignes	[18 ; **9**]	85,77% (15,01%) [158 ; 0]	14,58% (14,24%) [37 ; 9 ; **8**]	14,58% (14,24%) [37 ; 9 ; **8**]
LCP Info	[317 ; **96**]	61,42% (9,60%) [711 ; 0]	20,86% (10,05%) [517 ; 32 ; **28**]	20,89% (10,17%) [516 ; 32 ; **28**]
Pile et Face	[46 ; **15**]	43,62% (20,05%) [116 ; 0]	23,41% (19,72%) [60 ; 17 ; **14**]	23,41% (19,72%) [60 ; 17 ; **14**]
Top Questions	[119 ; **36**]	57,20% (13,31%) [346 ; 0]	28,23% (14,24%) [182 ; 26 ; **23**]	28,23% (14,24%) [182 ; 26 ; **23**]
Moyenne		**54,83% (14,92%)**	**20,92% (14,92%)**	**21,38% (14,97%)**

Table *B.1 – Résultats obtenus avec l'architecture de regroupement incrémental sur les collections de niveau* programme*, avec les approches de regroupement HAC$_{PLDA}$ et ILP$_{PLDA}$, avec application de l'approche de décomposition en composantes connexes. Pour chaque approche et chaque collection sont présentés : le DER$_{de\ collections}$, avec le DER$_{d'émissions}$ entre parenthèses, et le nombre de locuteurs détectés par le système (formalisme : [n^{bre} locuteurs total détecté par le système ; n^{bre} locuteurs récurrents détectés ;* **n^{bre} locuteurs détectés effectivement récurrents d'après les références***]).*

Les résultats en termes de DER$_{\text{de collections}}$ sont meilleurs, pour les deux méthodes de classification, avec l'approche de regroupement incrémental, comparés à ceux obtenus avec l'approche de regroupement global : la méthode de classification *CC+ILP*$_{PLDA}$ permet d'atteindre, en moyenne, 20,92% contre 23,29% pour l'approche de regroupement global (soit une diminution de 2,37% en absolu). La méthode *CC+HAC*$_{PLDA}$ permet quant à elle d'atteindre un DER$_{\text{de collections}}$ de 21,38%, en moyenne, contre 23,18% pour l'approche de regroupement global (soit une réduction de 1,8% en absolu). Contrairement au comportement observé avec l'approche de regroupement global, la méthode de classification *CC+ILP*$_{PLDA}$ semble plus efficace, avec l'approche de regroupement incrémental, que la méthode *CC+HAC*$_{PLDA}$. Les DER$_{\text{d'émissions}}$ moyens restent quant à eux relativement stables : il est strictement identique à celui obtenu avant la SRL de collections pour la méthode de classification *CC+ILP*$_{PLDA}$, et augmente de seulement 0,05% pour la méthode *CC+HAC*$_{PLDA}$.

Concernant la détection des locuteurs récurrents, seulement 54% de l'ensemble

des locuteurs récurrents sont correctement détectés par l'approche de regroupement incrémental, contre 58% pour l'approche de regroupement global.

Si l'on compare maintenant les résultats obtenus avec et sans application du procédé de *recyclage* des modèles de locuteurs (méthode de classification *CC+ILP*$_{PLDA}$ exclusivement, *cf.* partie 6.2), on constate une légère dégradation pour l'approche avec *recyclage* (*cf.* tableau B.2).

	CC. + ILP$_{PLDA}$ niveau collection	**CC. + ILP$_{PLDA}$ niveau collection + « Recyclage »**
Seuil β	-30	-30
Seuil δ	-30	-30
BFM Story	14,32% (10,35%) [678 ; 54 ; **43**]	15,09% (10,77%) [692 ; 56 ; **42**]
Planète Showbiz	47,35% (36,49%) [651 ; 16 ; **7**]	51,35% (37,10%) [675 ; 19 ; **6**]
Ça vous Regarde	21,90% (18,51%) [183 ; 8 ; **5**]	22,08% (18,51%) [186 ; 5 ; **3**]
Entre les Lignes	14,58% (14,24%) [37 ; 9 ; **8**]	15,34% (15,01%) [38 ; 9 ; **8**]
LCP Info	20,86% (10,05%) [517 ; 32 ; **28**]	22,03% (9,38%) [537 ; 38 ; **31**]
Pile et Face	23,41% (19,72%) [60 ; 17 ; **14**]	23,74% (20,05%) [61 ; 18 ; **14**]
Top Questions	28,23% (14,24%) [182 ; 26 ; **23**]	28,88% (13,96%) [187 ; 29 ; **25**]
Moyenne	**20,92% (14,92%)**	**21,87% (15,03%)**

Table *B.2 – Impact du procédé de* recyclage *en termes de DER$_{de\ collections}$ et $_{d'émissions}$ (entre parenthèses), ainsi qu'en termes de locuteurs détectés par le système (formalisme : [n^{bre} locuteurs total détecté par le système ; n^{bre} locuteurs récurrents détectés ;* **n^{bre} locuteurs détectés effectivement récurrents d'après les références***]), pour les collections d'émissions de niveau* Programme.

Le DER$_{\text{de collection}}$ augmente de 0,95%, en moyenne, avec l'approche implémentant le procédé de *recyclage*. Le DER$_{\text{d'émissions}}$ augmente également, mais plus modérément (+0,11%). La proportion de locuteurs récurrents correctement détectés par l'approche *CC+ILP*$_{PLDA}$*+Recyclage* diminue légèrement, mais reste également très proche : 52,8%. Ces pertes sont toutefois négligeables au regard du gain obtenu en termes de durée de traitement, et le DER$_{\text{de collections}}$ obtenu (21,87%) reste inférieur aux DER$_{\text{de collections}}$ enregistrés avec l'approche de regroupement global.

B.2. Collections temporelles

Les résultats obtenus en termes de DER sur les différentes collections temporelles, par les deux méthodes de classification $CC{+}ILP_{PLDA}$ et $CC{+}HAC_{PLDA}$, sont très proches, et même identiques pour les collections n°1, 2, 3, 4, 5 et 7 (*cf.* tableau B.3).

La différence en termes de $DER_{\text{de collections}}$ moyen entre les deux méthodes de classification n'est que de 0,02% : la méthode $CC{+}ILP_{PLDA}$ permet d'atteindre 20,69%, contre 20,71% pour la méthode $CC{+}HAC_{PLDA}$. Les $DER_{\text{d'émissions}}$ moyens sont quant à eux identiques. Si l'on compare ces résultats moyens à ceux obtenus avec l'approche de regroupement global, on constate qu'ils sont légèrement supérieurs, mais la différence est minime : la méthode $CC{+}ILP_{PLDA}$, en moyenne avec l'approche de regroupement global, donne un $DER_{\text{de collections}}$ de 20,66%, et la méthode $CC{+}HAC_{PLDA}$, 20,37%.

	n^{bre} locuteurs références	ILP_{PLDA} niveau émission	CC. + ILP_{PLDA} niveau collection	CC. + HAC_{PLDA} niveau collection
Seuil β		-	10	10
Seuil δ		20	10	40
Temporelle n°1	[58 ; **10**]	27,21% (11,71%) [89 ; 0]	13,51% (12,79%) [67 ; 11 ; **10**]	13,51% (12,79%) [67 ; 11 ; **10**]
Temporelle n°2	[217 ; **26**]	28,15% (13,19%) [263 ; 0]	21,74% (14,72%) [212 ; 25 ; **14**]	21,74% (14,72%) [213 ; 25 ; **14**]
Temporelle n°3	[140 ; **28**]	33,29% (12,91%) [170 ; 0]	15,51% (12,72%) [138 ; 23 ; **20**]	15,51% (12,72%) [138 ; 23 ; **20**]
Temporelle n°4	[669 ; **108**]	49,63% (13,79%) [986 ; 0]	19,29% (13,58%) [621 ; 66 ; **55**]	19,29% (13,58%) [620 ; 66 ; **55**]
Temporelle n°5	[247 ; **51**]	42,87% (13,31%) [402 ; 0]	20,94% (14,85%) [285 ; 33 ; **27**]	20,94% (14,85%) [285 ; 33 ; **27**]
Temporelle n°6	[205 ; **30**]	65,38% (24,66%) [309 ; 0]	30,57% (25,55%) [222 ; 23 ; **16**]	30,55% (25,46%) [223 ; 23 ; **16**]
Temporelle n°7	[262 ; **37**]	38,03% (11,88%) [329 ; 0]	14,26% (11,82%) [267 ; 28 ; **21**]	14,26% (11,82%) [267 ; 28 ; **21**]
Temporelle n°8	[461 ; **64**]	50,23% (19,14%) [682 ; 0]	26,94% (21,53%) [445 ; 55 ; **39**]	24,04% (21,52%) [442 ; 53 ; **39**]
Moyenne		**44,62% (14,99%)**	**20,69% (15,78%)**	**20,71% (15,78%)**

Table *B.3 – Résultats obtenus avec l'architecture de regroupement incrémental sur les collections* temporelles*, avec les approches de regroupement HAC_{PLDA} et ILP_{PLDA}, avec application de l'approche de décomposition en composantes connexes. Pour chaque approche et chaque collection sont présentés : le $DER_{\text{de collections}}$, avec le $DER_{\text{d'émissions}}$ entre parenthèses, et le nombre de locuteurs détectés par le système (formalisme : [n^{bre} locuteurs total détecté par le système ; n^{bre} locuteurs récurrents détectés ;* **n^{bre} locuteurs détectés effectivement récurrents d'après les références***]).*

Les résultats sont donc très proches, avec les collections temporelles, quelle que

soit l'approche de regroupement évaluée. Nous avions cependant établi le constat inverse avec les collections d'émissions de niveau *Programme*, où une différence significative pouvait être observée entre les résultats des approches de regroupement global et incrémental. Cette différence était d'ailleurs en faveur de l'approche de regroupement incrémental, contrairement à ce que nous pouvons observer avec les collections temporelles. Le nombre de locuteurs récurrents présent dans les collections temporelles est cependant moindre que dans les collections d'émissions. Notre pressentiment est que l'approche de regroupement incrémental permet une meilleure détection des locuteurs récurrents. Ce n'est pas flagrant avec les collections temporelles étant donné le faible nombre de locuteurs récurrents, en revanche, ça l'est plus avec les collections d'émissions. Nous aurions besoin des résultats sur la collection d'émissions de niveau *Thématique* pour confirmer cette hypothèse, cependant l'expérience est toujours en cours.

L'approche de regroupement incrémental ne permet pas de détecter autant de locuteurs récurrents que le fait l'approche de regroupement global. En moyenne sur les collections temporelles, l'approche de regroupement incrémental détecte correctement 62,5% de l'ensemble des locuteurs récurrents avec la méthode de classification $CC{+}ILP_{PLDA}$ (respectivement, 60,9% avec la méthode $CC{+}ILP_{PLDA}$). L'approche de regroupement global, sur les collections temporelles, permet de détecter 65% de l'ensemble des locuteurs récurrents.

La comparaison des résultats obtenus avec et sans application du procédé de *recyclage* des modèles de locuteurs (méthode de classification $CC{+}ILP_{PLDA}$ exclusivement, *cf.* partie 6.2), permet de constater une très légère amélioration des $\text{DER}_{\text{de collections}}$ et $\text{DER}_{\text{d'émissions}}$ (*cf.* tableau B.4). Les résultats restent proches, nous avions cependant observé le comportement inverse avec les collections d'émissions de niveau *Programme*. Le $\text{DER}_{\text{de collection}}$ diminue de 0,16%, en moyenne, avec l'approche implémentant le procédé de *recyclage*. Le $\text{DER}_{\text{d'émissions}}$ diminue quant à lui de 0,39%.

La proportion de locuteurs récurrents correctement détectés par l'approche $CC{+}ILP_{PLDA}{+}Recyclage$ augmente également légèrement (63,1%), mais reste inférieure à la proportion obtenue avec l'approche de regroupement global.

	CC. + ILP_{PLDA} niveau collection	**CC. + ILP_{PLDA} niveau collection + « Recyclage »**
Seuil β	10	10
Seuil δ	10	10
Temporelle n°1	13,51% (12,79%) [67 ; 11 ; **10**]	13,51% (12,79%) [67 ; 11 ; **10**]
Temporelle n°2	21,74% (14,72%) [212 ; 25 ; **14**]	21,90% (14,88%) [216 ; 26 ; **14**]
Temporelle n°3	15,51% (12,72%) [138 ; 23 ; **20**]	15,61% (12,81%) [139 ; 23 ; **20**]
Temporelle n°4	19,29% (13,58%) [621 ; 66 ; **55**]	18,96% (13,32%) [652 ; 75 ; **56**]
Temporelle n°5	20,94% (14,85%) [285 ; 33 ; **27**]	21,22% (14,31%) [299 ; 36 ; **25**]
Temporelle n°6	30,57% (25,55%) [222 ; 23 ; **16**]	31,27% (24,20%) [232 ; 25 ; **17**]
Temporelle n°7	14,26% (11,82%) [267 ; 28 ; **21**]	13,86% (11,67%) [270 ; 28 ; **22**]
Temporelle n°8	26,94% (21,53%) [445 ; 55 ; **39**]	26,50% (20,67%) [464 ; 58 ; **40**]
Moyenne	**20,69% (15,78%)**	**20,53% (15,39%)**

Table *B.4 – Impact du procédé de* recyclage *en termes de* $DER_{\text{de collections}}$ *et* $_{\text{d'émissions}}$ *(entre parenthèses), ainsi qu'en termes de locuteurs détectés par le système (formalisme : [*n^{bre} *locuteurs total détecté par le système ;* n^{bre} *locuteurs récurrents détectés ;* **n^{bre} locuteurs détectés effectivement récurrents d'après les références***]), pour les collections temporelles.*

B.3. Bilan

Le comportement de l'approche de regroupement incrémental diffère selon que les collections traitées soient des collections d'émissions ou des collections temporelles. Ce comportement est difficilement qualifiable. On observe un gain, en termes de $DER_{\text{de collections}}$, sur les collections d'émissions, et une relative stabilité sur les collections temporelles (comparé à l'approche de regroupement global). Le procédé de *recyclage* dégrade légèrement les $DER_{\text{de collections}}$ sur les collections d'émissions, et les améliore légèrement sur les collections temporelles (comparé à la version $CC{+}ILP_{PLDA}$ où le procédé de *recyclage* n'est pas implémenté). Une explication potentielle à ce phénomène peut être l'ordre dans lequel les enregistrements des collections sont traités, qui correspond à l'ordre chronologique. Or les collections temporelles ont été construites en respectant cet ordre chronologique, contrairement aux collections d'émissions.

ANNEXE C

Étude préliminaire sur le regroupement global

Le commencement de cette thèse a coïncidé avec l'arrivée au LIUM de Mickaël Rouvier. Son approche expérimentale de regroupement par Programmation Linéaire en Nombres Entiers, où les classes de locuteurs sont représentées par modélisation i-vector, présentait des caractéristiques intéressantes pour le traitement des collections d'enregistrements : modélisation plus robuste et regroupement plus rapide que la méthode à l'état de l'art basée sur la modélisation GMM et le regroupement hiérarchique ascendant (agglomératif). Fort des résultats présentés dans [Rouvier et Meignier, 2012], nos premiers travaux sur la SRL de collections ont consisté à comparer, sur plusieurs collections de tailles moindres que celles étudiées dans ce manuscrit, l'approche état de l'art et l'approche ILP_{Maha}.

Le système de SRL de collection mis en œuvre pour effectuer cette comparaison s'inspire des architectures présentées dans [Tran et al., 2011], détaillées dans la partie 3.2.2. Étant donné les conclusions des auteurs, nous avions opté pour une approche hybride, plus rapide que l'approche par concaténation globale, et plus robuste que l'approche incrémentale. Les données évaluées correspondaient alors à des collections constituées sur la base du corpus d'apprentissage fourni lors de la campagne d'évaluation ESTER 2. La configuration du système mis en place et les résultats de ces travaux préliminaires sont présentés en détail dans les sous-parties suivantes.

C.1. Approche de regroupement proposée

L'architecture de SRL de collections que nous avions proposée repose sur l'approche hybride de [Tran et al., 2011], et ne diffère de l'architecture de regroupement global présentée en chapitre 5 que par la configuration des approches de regroupement mises en œuvre. Dans cette version, illustrée en figure C.1, la dernière étape de regroupement du système de SRL d'émissions, pour le traitement local aux émissions, correspond à un regroupement agglomératif hiérarchique reposant sur le rapport de vraisemblance croisé (CLR) [Reynolds et al., 1998], et les classes de locuteurs sont modélisées par des modèles GMM. Compte tenu des performances obtenues par l'approche ILP_{Maha}, présentée dans [Rouvier et Meignier, 2012], nous nous sommes proposé de comparer les performances de cette approche de regroupement, pour le traitement global de la collection, avec l'approche de regroupement état de l'art HAC_{GMM}.

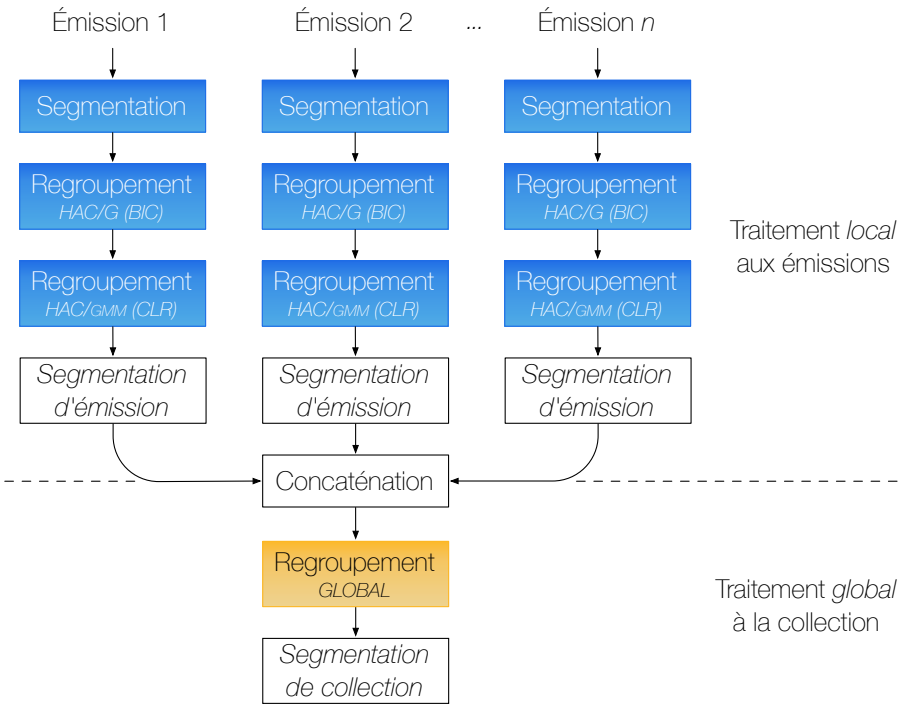

Figure *C.1 – Version de l'architecture de regroupement global mettant en œuvre un regroupement HAC_{GMM} pour le traitement local aux émissions.*

Approche HAC_{GMM} : Les classes de locuteurs sont représentées par des modèles GMM obtenus, pour chaque classe de la segmentation concaténée, à partir d'une adaptation MAP d'un GMM-UBM. Ce modèle du monde, indépendant du genre et de la brande de fréquence, est conçu avec 512 composantes gaussiennes. La paramétrisation utilisée correspond à 12 paramètres MFCC, plus

l'énergie, et sont normalisés par les méthodes de *features warping* et CMS. Afin de gagner du temps sur l'estimation des nouveaux modèles GMM résultant de la fusion de deux classes, lors du regroupement hiérarchique, la méthode employée par [Leeuwen, 2010] est appliquée : les nouveaux modèles GMM sont obtenus en fusionnant les accumulateurs statistiques des modèles GMM correspondants aux classes fusionnées (une seule itération de l'adaptation MAP est nécessaire). La mesure utilisée pour estimer la vraisemblance entre les modèles GMM est le rapport de vraisemblance croisé normalisé (NCLR). Afin d'accélérer la phase de calcul des mesures NCLR entre le nouveau modèle GMM résultant d'une fusion, et les autres, seules les 5 premières gaussiennes (top-5) sont considérées. Le modèle du monde a été entraîné sur les données de test distribuées durant la campagne d'évaluation française ESTER 1.

Approche ILP_{Maha} : L'étude présentée par [Rouvier et Meignier, 2012] montre que cette approche de regroupement permet d'obtenir de meilleurs résultats que l'approche hiérarchique lorsque les classes de locuteurs sont représentées par des modèles i-vector. Nous avons suivi la même « recette » . L'expression du problème de regroupement ILP est celle décrite dans la partie 2.3.4. Un modèle i-vector de dimension 60 a été extrait pour chacune des classes de locuteurs de la segmentation concaténée. La paramétrisation correspond à 19 paramètres MFCC + l'énergie, ainsi que leurs coefficients différentiels Δ et $\Delta\Delta$ respectifs. Le modèle du monde GMM-UBM, indépendant du genre et de la bande de fréquence et constitué de 1024 composantes gaussiennes, a été appris sur les données ESTER 1 à l'aide de la suite d'outils *Alize* [Bonastre et al., 2008] pour la reconnaissance du locuteur. La mesure de vraisemblance utilisée pour estimer la similarité entre les modèles i-vectors est la distance de Mahalanobis. Le programme d'optimisation linéaire utilisé pour résoudre le problème ILP est le l'outil de résolution *glpsol*, distribué librement dans la suite GPL *GNU GLPK*. La normalisation des modèles i-vector a été effectuée à l'aide des données d'apprentissage fournies lors de la campagne d'évaluation ESTER 1.

Les deux approches de regroupement global ont été configurées à l'aide d'un corpus de développement (*cf.* partie suivante) pour déterminer les seuils de décision optimaux. Concernant le traitement local aux émissions de la collection, le paramètre λ de la mesure BIC du premier regroupement local a été fixé à 3, et le seuil NCLR du deuxième regroupement local a été fixé à 0,97. Le seuil NCLR du regroupement global HAC_{GMM} a été fixé à 0,82, et la distance de Mahalanobis δ permettant d'atteindre les meilleurs $\text{DER}_{\text{de collections}}$ sur le corpus d'apprentissage, pour regroupement global ILP_{Maha}, a été fixée à 120.

C.2. Données expérimentales

Les données sélectionnées pour expérimenter ce système de SRL de collections, et comparer les deux approches de regroupement global, représentent l'intégralité du corpus d'apprentissage distribué durant la campagne d'évaluation ESTER 2. Ce corpus, dont la composition n'a pas été donnée dans le chapitre 4, est composé d'une centaine d'heures d'enregistrements radiophoniques d'émissions journalistiques, annotées manuellement, enregistrées entre 1999 et 2003.

Corpus	Radio (année)	Plage horaire	Nombre d'enregistrements	Durée (heures)	n^{bre} locuteurs
Col.dev	RFI *(2000)*	9:30 - 10:30	15	15	358 [206 : 48]
Col.1	RFI *(2000)*	11:30 - 12:30	15	15	298 [143 : 41]
Col.2	France Inter *(1999)*	19:00 - 19:20	5	2	66 [50 : 11]
Col.3	France Inter *(1999)*	7:00 - 8:00	10	10	235 [139 : 50]
Col.4	France Inter *(1999)*	8:00 - 9:00	10	10	181 [94 : 24]
Col.5	RFI *(2001)*	9:00 - 10:00	9	9	256 [165 : 45]
Col.6	RFI *(2001)*	10:00 - 11:00	9	9	244 [110 : 28]
Col.7	France Inter *(2002)*	19:00 - 20:00	5	5	151 [68 : 15]
Col.8	RFI *(2002)*	8:00 - 9:00	5	5	115 [88 : 20]
Col.9	RFI *(2002)*	0:00 - 1:00	5	5	113 [91 : 15]
Col.10	RFI *(2002)*	14:00 - 15:00	5	5	141 [93 : 21]
Col.11	RFI *(2002)*	20:00 - 21:00	5	5	166 [91 : 20]
Col.12	Africa *(2003)*	Toutes	13	5	130 [91 : 19]

Table *C.1 – Présentation des collections conçues à partir du corpus d'apprentissage ESTER 2, avec pour chaque sous-corpus, le nombre d'enregistrements qui les composent, leurs durées totale, et le n^{bre} de locuteurs [n^{bre} locuteurs* fiables *; n^{bre} locuteurs* fiables *récurrents].*

Ce corpus d'apprentissage a été divisé en 13 collections, sur lesquelles nous avons expérimenté notre système. La répartition des enregistrements au sein des collections a été influencée par l'identité de la radio émettrice, l'année, et la plage horaire de diffusion de chaque enregistrement. Cette répartition est présentée en détail dans le tableau C.1. La collection la plus volumineuse en termes de locuteurs a été utilisée comme corpus de développement, afin de déterminer les seuils et distances des approches de regroupement. Les 12 autres collections ont été considérées comme des corpus de tests.

C.3. Résultats et discussion

Les méthodes de regroupement global ILP_{Maha} et HAC_{GMM} ont été évaluées sur les 12 collections de test. Les taux d'erreur $DER_{d'émissions}$ et $DER_{de\ collections}$ sont

présentés dans le tableau C.2. Ce tableau présente également le nombre de classes initiales des problèmes de regroupement global, ainsi que la durée nécessaire à leurs traitements. Les durées de traitement sont exprimées en fraction du temps réel *(TR)* des enregistrements qui composent les collections. La moyenne des taux d'erreur a été pondérée en fonction de la durée des segments évalués dans chaque collection, et la moyenne des durées de traitement a, quant à elle, été pondérée par la durée des enregistrements de chaque collection.

Collection	$DER_{d'émissions}$		$DER_{de\ collections}$		n^{bre} de classes	Durée ($\times$TR)	
	ILP_{Maha}	HAC_{GMM}	ILP_{Maha}	HAC_{GMM}		ILP_{Maha}	HAC_{GMM}
Col.dev	8,50%	8,91%	15,06%	14,91%	670	*0,1799*	*3,2515*
Col.1	12,58%	12,29%	21,52%	19,97%	696	*0,1842*	*3,2171*
Col.2	3,23%	2,74%	8,14%	10,02%	74	*0,1836*	*0,2709*
Col.3	2,80%	2,94%	7,93%	10,75%	450	*0,1853*	*1,1854*
Col.4	3,79%	3,90%	5,16%	8,04%	323	*0,1639*	*0,8951*
Col.5	10,67%	10,66%	16,12%	16,65%	452	*0,1844*	*0,8657*
Col.6	15,09%	16,17%	20,98%	20,05%	440	*0,1858*	*0,8325*
Col.7	8,37%	8,26%	11,23%	11,59%	152	*0,1655*	*0,1878*
Col.8	6,16%	6,30%	8,92%	13,91%	200	*0,1740*	*0,3973*
Col.9	8,01%	5,40%	10,89%	8,67%	199	*0,1732*	*0,3502*
Col.10	8,71%	8,98%	15,25%	15,76%	219	*0,1835*	*0,4280*
Col.11	6,52%	6,64%	9,15%	9,43%	204	*0,1760*	*0,4090*
Col.12	15,77%	14,47%	27,30%	25,16%	216	*0,1814*	*0,4743*
Moyenne	8,81%	8,72%	14,32%	14,75%	330,4	*0,1785*	*1,1103*

Table *C.2 – $DER_{d'émissions}$ et $DER_{de\ collections}$ obtenus par les approches de regroupement ILP_{Maha} et HAC_{GMM} sur les collections confectionnées à partir du corpus d'apprentissage ESTER 2.*

Les taux d'erreur $DER_{d'émissions}$ et $DER_{de\ collections}$ sont relativement similaires entre les deux approches de regroupement global. La collection *Col.9* présente toutefois un caractère anormal : la différence entre les $DER_{d'émissions}$ des deux approches de regroupement est de 2,61% en absolu, en faveur du regroupement HAC_{GMM}. Cette différence est inversée entre les $DER_{de\ collections}$, avec 4,99% en absolu en faveur du regroupement ILP_{Maha}. En moyenne, la différence entre les deux approches de regroupement global n'est que de 0,09% en termes de $DER_{d'émissions}$ (en faveur de l'approche HAC_{GMM}), et 0,43% en termes de $DER_{de\ collections}$ (en faveur de l'approche ILP_{Maha}).

L'approche ILP_{Maha} se démarque essentiellement de l'approche HAC_{GMM} par la rapidité avec laquelle les collections sont traitées. Les durées de traitement présentées dans la tableau C.2 prennent en compte la concaténation des segmentations locales, l'extraction des paramètres acoustiques, la modélisation des classes, le processus de regroupement et la génération de la segmentation finale. Les durées de traitement sont très stables avec l'approche ILP_{Maha}, alors qu'elles dépendent manifestement du

nombre de classes initiales pour l'approche HAC_{GMM}. Pourtant, le regroupement hiérarchique a été optimisé de manière à en minimiser sa complexité : le modèle GMM d'une classe issue d'un regroupement correspond à la fusion des accumulateurs statistiques des modèles GMM des classes regroupées, et ses mesures de vraisemblance avec les autres modèles ne sont calculées que sur les 5 premières gaussiennes. La collection *Col.1*, qui présente le plus grand nombre de classes initiales, est traitée en *0,18* $\times$ le temps réel de la collection par l'approche ILP_{Maha}, contre *3,22* pour l'approche HAC_{GMM}. La collection *Col.2*, qui présente le plus faible nombre de classes initiales, est également traitée plus rapidement par l'approche ILP_{Maha} (*0,18 vs. 0,27* $\times TR$).

Cette étude préliminaire, publiée dans [Dupuy et al., 2012b,a], a permis d'apprécier l'intérêt de l'approche ILP_{Maha} par rapport à l'approche état de l'art HAC_{GMM} pour le traitement des collections : les taux d'erreurs sont sensiblement les mêmes, globalement supérieurs en termes de $\text{DER}_{\text{de collections}}$, et les collections sont traitées beaucoup plus rapidement.

ANNEXE D

Étude préliminaire sur le regroupement incrémental

Les premiers travaux que nous avons réalisés sur l'architecture de regroupement incrémental ont été menés dans le cadre du projet européen EUMSSI, pour lequel nous nous sommes posé la question du traitement des collections dynamiques, dont le volume augmenterait au cours du temps. La plupart des approches présentées dans le chapitre 6 ont d'abord été expérimentées dans cette étude préliminaire. Les données utilisées pour expérimenter nos approches, dans cette étude, correspondent à un sous-ensemble des données REPERE.

D.1. Approche de regroupement proposée

Le système de SRL de collections mis en place correspond à celui présenté en partie 6.3, où nous avons proposé de démarrer le procédé de regroupement incrémental à partir d'une collection initiale déjà traitée par regroupement global (*cf.* figure D.1). L'objectif de cette collection initiale préalablement traitée était double : d'une part, simuler une collection dont le volume augmente. Les enregistrements ajoutés à la collection initiale étant traités incrémentalement. D'autre part, observer l'influence de cette collection initiale sur les résultats obtenus au terme de la dernière itération de l'approche de regroupement incrémental. La constitution des collections initiales, dénommées *Boot.1* et *Boot.2*, et l'ordre dans lequel les enregistrements restants sont itérativement traités, est fonction de l'ordre chronologique de diffusion des enregistrements.

Dans cette étude préliminaire, les différents enregistrements composant les col-

lections sont d'abord traités indépendamment avec un système de SRL d'émissions proche de celui présenté pour la rédaction de ce manuscrit de thèse. La dernière étape de regroupement de ce système de SRL d'émissions repose sur l'approche de décomposition en composantes connexes et sur le regroupement combinatoire ILP (pour traiter les composantes connexes complexes (*cf.* partie 5.2.3)). Les classes de locuteurs issues du regroupement BIC sont représentées par des modèles i-vectors de dimension 50, calculés à partir d'un GMM-UBM de 256 composantes gaussiennes et normalisés par 5 itérations de l'algorithme EFR. La mesure employée pour estimer la similarité entre les modèles i-vector correspond à la distance de *Mahalanobis*.

La méthode de regroupement pour le niveau *collection* de l'architecture, qui est employée pour le traitement de la collection initiale, ainsi que pour chaque itération du procédé de regroupement incrémental, est identique en tout point à celle du niveau *émission* si ce n'est que le modèle du monde est remplacé par un GMM-UBM à 1024 composantes, et que la paramétrisation diffère légèrement.

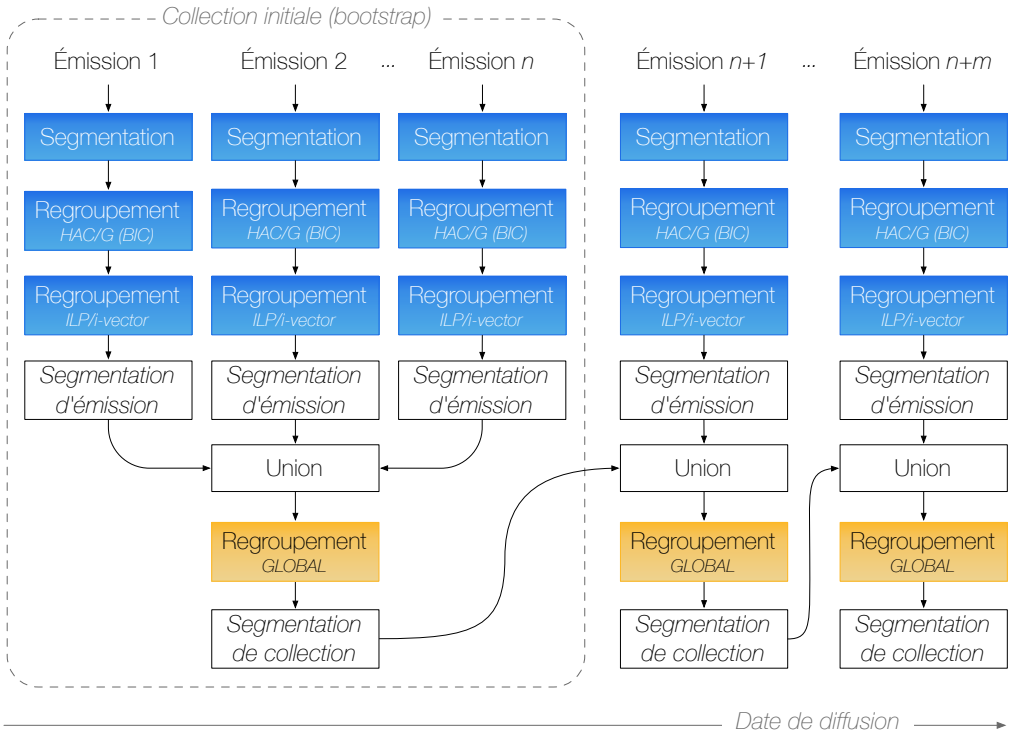

Figure *D.1 – Architecture par regroupement incrémental avec collection* initiale *pour la SRL de collections.*

Le procédé de *recyclage* des modèles de locuteurs avait déjà été implémenté lors de cette étude préliminaire : plutôt que d'apprendre un modèle sur l'ensemble des données d'une nouvelle classe, issue d'un regroupement, nous considérons le modèle de locuteur de la classe la plus centrale de ce regroupement pour le représenter.

D.2. Données expérimentales

Les données sélectionnées pour procéder aux expériences correspondent à tous les enregistrements des chaînes BFMTV et LCP distribués dans les corpus d'apprentissage, de développement et de test des campagnes d'évaluation françaises ETAPE et REPERE 2013. Nous avons également utilisé les données du corpus d'apprentissage fournies pour la campagne REPERE 2014. Cette collection est finalement constituée de 310 enregistrements audiovisuels, enregistrés entre septembre 2010 et octobre 2012. La durée totale de cette collection est de 142 heures, mais seulement 67 heures sont annotées à des fins d'évaluation.

Émission	Ensemble des données		Boot.1		Boot.2	
	n^{bre} enr.	Durée	n^{bre} enr.	Durée	n^{bre} enr.	Durée
BFMStory	37	21:21	7	05:01	19	10:36
CultureEtVous	54	01:42	0	-	0	-
PlaneteShowbiz	73	02:24	1	02:07	35	01:10
CaVousRegarde	20	09:02	9	04:29	12	05:21
EntreLesLignes	24	08:20	10	04:05	16	06:14
LCPInfo	35	08:37	1	00:10	10	01:57
PileEtFace	32	08:15	12	04:07	19	04:52
TopQuestions	35	07:41	9	02:03	15	03:26
Total	310	67:22	49	22:52	126	33:36

Table *D.1 – Nombre d'enregistrements et durées (h:min) pour l'ensemble des données évaluées ainsi que pour les deux collections initiales Boot.1 et Boot.2.*

Les 49 (respectivement, 126) premiers enregistrements de l'ensemble des données ont été sélectionnés pour former la collection initiale *Boot.1* (respectivement, *Boot.2*), et les enregistrements restants sont itérativement ajoutés les uns après les autres. *Boot.1* représente environ le tiers de l'ensemble des données, en termes de durée, et *Boot.2*, la moitié.

D.3. Résultats et discussion

Les $DER_{\text{de collections}}$ et $DER_{\text{d'émissions}}$ obtenus pour le regroupement global des collections initiales, et lors de chaque itération du regroupement incrémental, sont présentés en figure D.2. Les lignes horizontales parcourant le graphique en largeur représentent les $DER_{\text{de collections}}$ et $DER_{\text{d'émissions}}$ obtenus avec un regroupement global sur l'ensemble des 310 enregistrements. Le $DER_{\text{de collections}}$ obtenu par regroupement global sur l'ensemble des données est élevé (26,91%), comparé à ceux des collections initiales (20,28% pour *Boot.1* et 23,03% pour *Boot.2*). Les $DER_{\text{de collections}}$

obtenus avec l'approche incrémentale, après que le dernier enregistrement ait été traité, sont cependant très proches : 27,45% pour l'expérience démarrant avec la collection initiale *Boot.1*, et 25,97% pour celle démarrant avec *Boot.2*. En termes de $\text{DER}_{\text{d'émissions}}$, les résultats obtenus avec l'approche de regroupement incrémental sont quelque peu supérieurs au $\text{DER}_{\text{d'émissions}}$ obtenu par regroupement global sur les 310 enregistrements : 17,02% pour l'expérience démarrant avec *Boot.1* et 16,68% pour celle démarrant avec *Boot.2*, contre 15,11% pour le regroupement global sur l'ensemble des données.

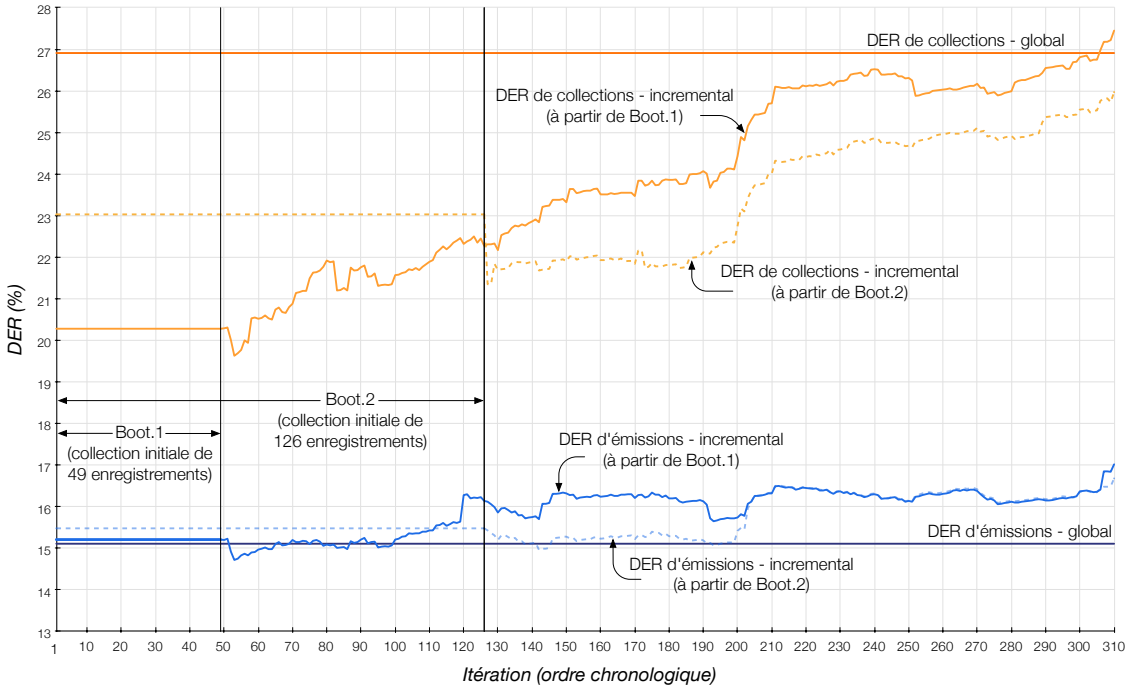

Figure *D.2 – $\text{DER}_{\text{de collections}}$ et $\text{DER}_{\text{d'émissions}}$ pour chaque itération du procédé de regroupement incrémental démarré à partir des collections initiales Boot.1 et Boot.2, ainsi que les $\text{DER}_{\text{de collections}}$ et $\text{DER}_{\text{d'émissions}}$ obtenus sur l'ensemble des données avec une approche de regroupement global (les enregistrements sont traités simultanément).*

La brutale augmentation des $\text{DER}_{\text{de collections}}$ entre les itérations 200 et 210 pourrait s'expliquer par les *ruptures* dans la fréquence d'acquisition des enregistrements : 24 jours séparent les enregistrements des itérations 201 et 202. Nous n'observons cependant pas un comportement similaire avec les autres *ruptures* dans la fréquence d'acquisition des enregistrements, et en retirant les enregistrements correspondant aux itérations 200 à 210, nous obtenons finalement des résultats globalement meilleurs et plus *linéaires*. Cela ne peut donc constituer qu'une explication partielle.

La collection initiale *Boot.1* représente 34% de la totalité des données audiovisuelles (respectivement, 50% pour *Boot.2*). 55,8% des locuteurs présents dans

Boot.1 sont également présents dans les enregistrements traités par l'approche incrémentale (respectivement, 58,9% pour *Boot.2*). 12 locuteurs sont particulièrement présents (principalement des présentateurs et des journalistes). Le temps de parole de ces 12 locuteurs représente environ 30% de la collection initiale *Boot.1*, et environ 20% de la collection initiale *Boot.2*. Au regard des résultats obtenus,

Cette étude préliminaire, publiée dans [Dupuy et al., 2014c], synthétise nos premiers travaux sur les architectures de regroupement incrémental. Nous avons pu observer, sans vraiment pouvoir en tirer de conclusions, que ce procédé ne dégrade pas forcément les résultats en termes de $\text{DER}_{\text{de collections}}$, comparé à un regroupement global. La taille de la collection initiale semble influencer les résultats, nous ne sommes cependant pas en mesure d'en affirmer la raison, car il est difficile de mener une analyse précise du fait de la quantité de données, mais également, de l'implémentation du procédé de *recyclage* des modèles de locuteurs entre les itérations.

Acronymes

AFCP	*Association Francophone de la Communication Parlée*
ANR	*Agence Nationale de la Recherche*
CE	*Cross Entropy*
BIC	*Bayesian Information Criterion*
CLR	*Cross Likelihood Ratio*
CMS	*Cepstral Mean Substraction*
DER	*Diarization Error Rate*
DCT	*Discrete Cosine Transform*
ELDA	*Evaluations and Language resources Distribution Agency*
ESTER	*Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques*
EUMSSI	*Event Understanding through Multimodal Social Stream Interpretation*
DGA	*Direction Générale de L'Armement*
ETAPE	*Évaluations en Traitement Automatique de la ParolE*
FFT	*Fast Fourier Transform*
GLR	*Generalized Likelihood Ratio*
GMM	*Gaussian Mixture Model*
HAC	*Hierarchical Agglomerative Clustering*
HMM	*Hidden Markov Model*
ILP	*Integer Linear Programming*
KL(2)	*Divergence de Kullback-Leibler*
LFCC	*Linear Frequency Cepstral Coefficients*
LPCC	*Linear Predictive Cepstral Coefficients*
LNE	*Laboratoire National de métrologie et d'Éssais*
MFCC	*Mel-Frequency Cepstral Coefficients*
MLLR	*Maximum Likelihood Linear Regression*
NIST	*National Institute of Standards and Technology*
PLNE	*Programmation Linéaire en Nombres Entiers*
PLP	*Perceptual Linear Prediction*
REPERE	*REconnaissance de PERsonnes dans des Émissions audiovisuelles*
RAP	*Reconnaissance Automatique de la Parole*
SAD	*Speech Activity Detection*

SODA *reconnaiSsance de persOnnes pour Débats et journAux télévisés*
SRL *Segmentation et Regroupement en Locuteurs*
UEM *Unpartitioned Evaluation Map*

Liste des figures

Liste des tableaux

Références bibliographiques

X. Anguera et J. Hernando, "Evolutive Speaker Segmentation Using a Repository System," in *Interspeech*, Jeju, Korea, September 2004.

B. S. Atal et S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

J. Baker, "The DRAGON System - An Overview," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 24–29, 1975.

C. Barras, X. Zhu, S. Meignier, et J. Gauvain, "Multi-stage Speaker Diarization of Broadcast News," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1505–1512, September 2006.

C. Barras, X. Zhu, S. Meignier, et J.-L. Gauvain, "Improving Speaker Diarization," in *RT-04F workshop*, 2004.

L. E. Baum et T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *The annals of mathematical statistics*, pp. 1554–1563, 1966.

B. Bigot, J. Pinquier, I. Ferrané, et R. André-Obrecht, "Detecting Individual Role Using Features Extracted from Speaker Diarization Results," *Multimedia Tools and Applications*, vol. 60, no. 2, pp. 347–369, septembre 2012.

J.-F. Bonastre, S. Meignier, et T. Merlin, "Speaker Detection Using Multi-speaker Audio Files for Both Enrollment and Test," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 2. IEEE, 2003, pp. II–77.

J.-F. Bonastre, F. Bimbot, L.-J. Boë, J. P. Campbell, D. A. Reynolds, et I. Magrin-Chagnolleau, "Person Authentication by Voice: a Need for Caution," in *INTERSPEECH*, 2003.

J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. W. Evans, B. G. Fauve, et J. S. Mason, "ALIZE/spkdet: a State-of-the-art Open Source Software for Speaker Recognition," in *Odyssey*, 2008, p. 20.

H. Bourlard, M. Ferras, N. Pappas, A. Popescu-Belis, S. Renals, F. McInnes, P. Bell, et M. Guillemot, "Processing and Linking Audio Events in Large Multimedia Archives: The EU inEvent Project," in *Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France, August 2013.

P.-M. Bousquet, D. Matrouf, et J.-F. Bonastre, "Intersession Compensation and Scoring Methods in the I-vectors Space for Speaker Recognition," in *Proceedings of Interspeech*, Florence, Italia, 2011.

P.-M. Bousquet, A. Larcher, D. Matrouf, J.-F. Bonastre, et O. Plchot, "Variance-spectra Based Normalization for i-vector Standard and Probabilistic Linear Discriminant Analysis," in *Odyssey: The Speaker and Language Recognition Workshop, Singapore, Singapore*, 2012, pp. 157–164.

H. Bredin et J. Poignant, "Integer Linear Programming for Speaker Diarization and Cross-modal Identification in TV Broadcast," in *Proceedings of Interspeech*, Lyon, France, 2013.

H. Bredin, A. Laurent, A. Sarkar, V.-B. Le, S. Rosset, et C. Barras, "Person Instance Graphs for Named Speaker Identification in TV Broadcast," in *Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop*, Joensuu, Finland, 2014.

N. Brümmer et E. De Villiers, "The Speaker Partitioning Problem," in *Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.

L. Burget, P. Matejka, P. Schwarz, O. Glembek, et J. Cernocky, "Analysis of Feature Extraction and Channel Compensation in a GMM Speaker Recognition System," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1979–1986, 2007.

J. Campbell, W. Shen, W. Campbell, R. Schwartz, J.-F. Bonastre, et D. Matrouf, "Forensic Speaker Recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, 2009.

W. M. Campbell et E. Singer, "Query-by-Example using Speaker Content Graphs," in *Interspeech*, 2012.

M. Carey et E. S. Parris, "Speaker Verification Using Connected Words," *Proceedings of Institute of Acoustics*, vol. 14, pp. p95–p100, 1992.

S. Chen et P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*. Virginia, USA, 1998, p. 8.

S. Cook, "The P Versus NP Problem," *The millennium prize problems*, p. 86, 2006.

N. Dehak, "Discriminative and Generative Approaches for Long-and Short-term Speaker Characteristics Modeling : Application to Speaker Verification," in *Thèse de doctorat*. École de Technologie Supérieure de Montréal (ETS) et Centre de Recherche en Informatique de Montréal (CRIM), 2009.

N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, et P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

P. Delacourt et C. J. Wellekens, "DISTBIC: A Speaker-based Segmentation for Audio Data Indexing," *Speech communication*, vol. 32, no. 1, pp. 111–126, 2000.

P. Delacourt, D. Kryze, et C. J. Wellekens, "Detection of Speaker Changes in an Audio Document," in *Sixth European Conference on Speech Communication and Technology*, 1999.

A. P. Dempster, N. M. Laird, et D. B. Rubin, "Maximum Likelihood From Incomplete Data via The EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

G. Dupuy, M. Rouvier, S. Meignier, et Y. Estève, "Segmentation et Regroupement en Locuteurs d'une collection de documents audio," in *Proceedings of 29e Journées d'Études sur la Parole (JEP'12)*, Grenoble, France, June 2012.

——, "I-vectors and ILP Clustering Adapted to Cross-show Speaker Diarization," in *Proceedings of Interspeech'12*, Portland, Oregon (USA), September 2012.

G. Dupuy, S. Meignier, P. Deléglise, et Y. Estève, "Recent Improvements on ILP-based Clustering for Broadcast News Speaker Diarization," in *Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop*, Joensuu, Finland, June 2014.

G. Dupuy, S. Meignier, et Y. Estève, "Segmentation et Regroupement en Locuteur pour le traitement incrémental des collections volumineuses," in *Proceedings of 30e Journées d'Études sur la Parole (JEP'14)*, Le Mans, France, June 2014.

——, "Is Incremental Cross-Show Speaker Diarization Efficient For Processing Large Volumes of Data?" in *Proceedings of Interspeech'14*, Singapore, September 2014.

M. Ferràs et H. Bourlard, "Speaker Diarization and Linking of Large Corpora," in *Proceedings of IEEE Workshop on Spoken Language Technology*, Miami, Florida (USA), December 2012.

C. Fredouille et N. Evans, "The LIA RT'07 Speaker Diarization System," in *Multimodal Technologies for Perception of Humans*. Springer, 2008, pp. 520–532.

S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 2, pp. 254–272, 1981.

O. Galibert, "Methodologies for the Evaluation of Speaker Diarization and Automatic Speech Recognition in the Presence of Overlapping Speech," in *Proceedings of Interspeech'13*, Lyon, France, August 2013.

O. Galibert et J. Kahn, "The First Official REPERE Evaluation," in *Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France, August 2013.

S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, et G. Gravier, "The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News," in *Proceedings of Eurospeech'05*, Lisbon, Portugal, September 2005.

S. Galliano, G. Gravier, et L. Chaubard, "The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts," in *Proceedings of Interspeech'09*, Brighton, UK, September 2009.

D. Garcia-Romero et C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Interspeech*, 2011, pp. 249–252.

J.-L. Gauvain et C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE transactions on Speech and audio processing*, vol. 2, no. 2, pp. 291–298, 1994.

J. T. Geiger, F. Wallhoff, et G. Rigoll, "GMM-UBM Based Open-Set Online Speaker Diarization," in *Interspeech*, 2010, pp. 2330–2333.

H. Ghaemmaghami, D. Dean, et S. Sridha, "Speaker Attribution of Australian Broadcast News Data," in *Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France, August 2013.

H. Gish, M.-H. Siu, et R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1991, pp. 873–876.

O. Glembek, L. Burget, N. Dehak, N. Brummer, et P. Kenny, "Comparison of Scoring Methods Used in Speaker Recognition With Joint Factor Analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

G. Gravier, G. Adda, N. Paulsson, M. Carré, A. Giraudel, et O. Galibert, "The ETAPE Corpus for the Evaluation of Speech-based TV Content Processing in the French Language," in *Proceedings of LREC'12*, Istanbul, Turkey, May 2012.

V. Gupta, G. Boulianne, P. Kenny, P. Ouellet, et P. Dumouchel, "Speaker Diarization of French Broadcast News," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 4365–4368.

Gurobi Optimization, Inc., "Gurobi Optimizer Reference Manual," 2015. [Online]. Available: http://www.gurobi.com

Hagai Aronowitz, "Interspeech 2014 Tutorial: Recent Advances in Speaker Diarization," 2014. [Online]. Available: https://sites.google.com/site/aronowitzh/

F. Jelinek, "Speech Recognition by Statistical Methods," *Proceedings of the IEEE*, vol. 64, pp. 532–556, 1976.

Y. Jiang, K.-A. Lee, Z. Tang, B. Ma, A. Larcher, et H. Li, "PLDA Modeling in i-vector and Supervector Space for Speaker Verification," in *INTERSPEECH*, 2012.

S. Johnson, "Who Spoke When? - Automatic Segmentation and Clustering for Determining Speaker Turns," in *EuroSpeech*, 1999.

S. Johnson et P. Woodland, "Speaker Clustering Using Direct Maximisation of the MLLR-Adapted Likelihood," in *International Conference on Spoken Language Processing (ICSLP)*, 1998.

Z. N. Karam et W. M. Campbell, "Graph Embedding for Speaker Recognition," in *Graph Embedding for Pattern Analysis*. Springer, 2013, pp. 229–260.

P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Odyssey*, 2010, p. 14.

P. Kenny, M. Mihoubi, et P. Dumouchel, "New MAP Estimators for Speaker Recognition," in *Proceedings of Interspeech'03*, 2003.

P. Kenny, G. Boulianne, P. Ouellet, et P. Dumouchel, "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1435–1447, 2007.

P. Kenny, P. Ouellet, N. Dehak, V. Gupta, et P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, et M. Contolini, "Eigenvoices for Speaker Adaptation," in *ICSLP*, vol. 98, 1998, pp. 1774–1777.

G. Lanc et W. Williams, "A general theory of classificatory sorting strategies," *Computer Journal*, vol. 9, pp. 373–380, 1967.

V. B. Le, O. Mella, et D. Fohr, "Speaker Diarization Using Normalized Cross Likelihood Ratio," in *Proceedings of Interspeech'07*, vol. 7, 2007, pp. 1869–1872.

D. A. Leeuwen, "Speaker Linking in Large Data Sets," in *Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.

C. J. Leggetter et P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

K. Markov et S. Nakamura, "Never-Ending Learning System for On-Line Speaker Diarization," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2007, pp. 699–704.

P. Matejka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, et J. Cernocky, "Full-covariance UBM and Heavy-tailed PLDA in i-vector Speaker Verification," in *Proceedings of Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4828–4831.

D. Matrouf, N. Scheffer, B. G. Fauve, et J.-F. Bonastre, "A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification," in *Proceedings of Interspeech'07*, 2007.

D. Matrouf, J.-F. Bonastre, et S. E. Mezaache, "Factor Analysis Multi-session Training Constraint in Session Compensation for Speaker Verification," in *Proceedings of Interspeech'08*, 2008.

S. Meignier, "Indexation en locuteurs de documents sonores : segmentation d'un document et appariemment d'une collection," in *Thèse de doctorat*. Laboratoire d'Informatique d'Avignon, 2002.

——, "Détection et Identification des Locuteurs des Émissions Radiophoniques et Télévisées," in *Habilitation à Diriger des Recherches (HDR)*. LIUM, Université du Maine (FRANCE), 2015.

S. Meignier et T. Merlin, "LIUM SpkDiarization: an open-source toolkit for diarization," in *CMU SPUD Workshop*, Dallas, Texas (USA), 2009.

S. Meignier, J.-F. Bonastre, et S. Igounet, "E-HMM Approach for Learning and Adapting Sound Models for Speaker Indexing," in *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, 2001.

S. Meignier, J.-F. Bonastre, et I. Magrin-Chagnolleau, "Speaker Utterances Tying Among Speaker Segmented Audio Documents Using Hierarchical Classification: Towards Speaker Indexing of Audio Databases," in *INTERSPEECH*, 2002.

S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, et L. Besacier, "Step-by-step and Integrated Approaches in Broadcast News Speaker Diarization," *Computer Speech & Language*, vol. 20, no. 2, pp. 303–330, 2006.

H. Ning, M. Liu, H. Tang, et T. S. Huang, "A Spectral Clustering Approach to Speaker Diarization," in *Proceedings of Interspeech'06,*, 2006.

NIST, "The 2002 NIST Speaker Recognition Evaluation Plan," April 2002. [Online]. Available: http://www.nist.gov/speech/tests/spk/2002/2002-spkrec-evalplan-v60.pdf

——, "The NIST Rich Transcription Spring 2003 (RT-03S) Evaluation Plan," March 2003. [Online]. Available: http://www.nist.gov/speech/tests/rt/2003-spring/docs/rt03-spring-eval-plan-v4.pdf

——, "Fall 2004 Rich Transcription (RT-04F) Evaluation Plan," October 2004. [Online]. Available: http://www.nist.gov/speech/tests/rt/2004-fall/docs/rt04f-eval-plan-v14.pdf

——, "Spring 2006 (RT-06S) Rich Transcription Meeting Recognition Evaluation Plan," March 2006. [Online]. Available: http://www.nist.gov/speech/tests/rt/2006-spring/docs/rt06s-meeting-eval-plan-V2.pdf

——, "The NIST Year 2010 Speaker Recognition Evaluation Plan," March 2010. [Online]. Available: http://www.nist.gov/speech/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf

A. V. Oppenheim et R. Schafer, *Digital Signal Processing*, ser. Prentice-Hall international editions, 1975.

P. Ouellet, G. Boulianne, et P. Kenny, "Flavors of Gaussian Warping," in *Proceedings of Interspeech'05 - Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, January 2005.

J. Pelecanos et S. Sridharan, "Feature Warping for Robust Speaker Verification," in *Proceedings of Odyssey 2001: The Speaker Recognition Workshop*, Crete, Greece, June 2001.

S. J. Prince et J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

L. R. Rabiner et B.-H. Juang, *Fundamentals of Speech Recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

L. Rabiner et R. Schafer, *Digital Processing of Speech Signals*, ser. Prentice-Hall signal processing series, 1978.

D. A. Reynolds et P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization," in *Proceedings of ICASSP'05*, Philadelphia, Pennsylvania (USA), March 2005.

D. A. Reynolds, "A Gaussian Mixture Modeling Approach to Text-independent Speaker Identification," in *Thèse de doctorat*. Georgia Institute of Technology, 1992.

D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O'Leary, J. McLaughlin, et M. A. Zissman, "Blind Clustering of Speech Utterances Based on Speaker and Language Characteristics," in *International Conference on Spoken Language Processing (ICSLP)*, 1998.

D. A. Reynolds, R. B. Dunn, et J. McLaughlin, "The Lincoln Speaker Recognition System: NIST eval2000," in *Proceedings of Interspeech*, 2000, pp. 470–473.

D. A. Reynolds, T. F. Quatieri, et R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

R. C. Rose et D. A. Reynolds, "Text Independent Speaker Identification Using Automatic Acoustic Segmentation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1990, pp. 293–296.

M. Rouvier et S. Meignier, "A Global Optimization Framework for Speaker Diarization," in *Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop*, Singapore, 2012.

M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, et S. Meignier, "An Open-source State-of-the-art Toolbox for Broadcast News Diarization," in *Proceedings of Interspeech'13*, Lyon, France, August 2013.

G. Schwarz *et al.*, "Estimating the Dimension of a Model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, et J. Glass, "Exploiting Intra-Conversation Variability for Speaker Diarization," in *Proceedings of Interspeech*, Florence, Italy, 2011.

S. Shum, N. Dehak, et J. Glass, "On the Use of Spectral and Iterative Methods for Speaker Diarization," in *Proceedings of Interspeech'12*, Portland, Oregon (USA), September 2012.

S. H. Shum, W. M. Campbell, et D. A. Reynolds, "Large-scale Community Detection on Speaker Content Graphs," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7716–7720.

M. A. Siegler, U. Jain, B. Raj, et R. M. Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," in *Proceedings of the DARPA Broadcast News Workshop*, 1997, p. 11.

M.-H. Siu, G. Yu, et H. Gish, "An Unsupervised, Sequential Learning Algorithm for the Segmentation of Speech Waveforms with Multiple Speakers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1992, pp. 189–192.

A. Solomonoff, A. Mielke, M. Schmidt, et H. Gish, "Clustering Speakers by Their Voices," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2. IEEE, 1998, pp. 757–760.

V.-A. Tran, V. B. Le, C. Barras, et L. Lamel, "Comparing Multi-stage Approaches for Cross-Show Speaker Diarization," in *Proceedings of Interspeech'11*, Florence, Italie, August 2011.

S. Tranter et D. A. Reynolds, "Speaker Diarisation for Broadcast News," in *Odyssey 2004: The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004.

S. Tranter, M. Gales, R. Sinha, S. Umesh, et P. Woodland, "The Development of the Cambridge University RT-04 Diarisation System," in *The Fall 2004 Rich Transcription Workshop (RT-04)*, 2004.

D. A. van Leeuwen et N. Brümmer, "Constrained Speaker Linking," *arXiv preprint arXiv:1403.7084*, 2014.

A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

J. H. Ward Jr, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

Q. Yang, Q. Jin, et T. Schultz, "Investigation of Cross-show Speaker Diarization," in *Proceedings of Interspeech'11*, Florence, Italie, August 2011.

M. Zelenák, H. Schulz, et J. Hernando, "Speaker Diarization of Broadcast News in Albayzin 2010 Evaluation Campaign," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2012, no. 1, pp. 1–9, 2012.

Liste des publications personelles

- **Conférences d'audience internationale avec comité de relecture**

[Dupuy 2014b] Dupuy G., Meignier S. et Estève Y., Is Incremental Cross-Show Speaker Diarization Efficient For Processing Large Volumes of Data?, dans Proceedings of Interspeech'14, Singapor, September 2014b.

[Rouvier 2013] Rouvier M., Dupuy G., Gay P., Khoury E., Merlin T. et Meignier S., An Open-source State-of-the-art Toolbox for Broadcast News Diarization, dans Proceedings of Interspeech'13, Lyon, France, August 2013.

[Dupuy 2012a] Dupuy G., Rouvier M., Meignier S. et Estève Y., I-vectors and ILP Clustering Adapted to Cross-Show Speaker Diarization, dans Proceedings of Interspeech'12, Portland, Oregon (USA), September 2012a.

- **Workshops d'audience internationale avec comité de relecture**

[Dupuy 2014a] Dupuy G., Meignier S., Deléglise P. et Estève Y., Recent Improvements on ILP-based Clustering for Broadcast News Speaker Diarization, dans Odyssey 2014: The Speaker and Language Recognition Workshop, Joensuu, Finland, June 2014a.

[Gay 2014] Gay P., Dupuy G., Odobez J.-M., Meignier S. et Deléglise P., Comparison of Two Methods for Unsupervised Person Identification in TV Shows, dans 12th International Workshop on Content-Based Multimedia Indexing (CBMI'14), Klagenfurt, Austria, June 2014.

[Lailler 2013] Lailler C., Dupuy G., Rouvier M. et Meignier S., Semi-Supervised and Unsupervised Data Extraction Targeting Speakers: From Speaker Roles to Fame?,

dans Proceedings of Interspeech'13 satellite workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 2013.

- **Conférences d'audience nationale avec comité de relecture**

[Dupuy 2014c] Dupuy G., Meignier S. et Estève Y., Segmentation et Regroupement en Locuteur pour le traitement incrémental des collections volumineuses, dans 30e Journées d'Études sur la Parole (JEP'14), Le Mans, France, June 2014b.

[Dupuy 2012b] Dupuy G., Rouvier M., Meignier S. et Estève Y., Segmentation et Regroupement en Locuteurs d'une collection de documents audio, dans 29e Journées d'Études sur la Parole (JEP'12), Grenoble, France, June 2012b.